MASTER EN SCIENCES MATHÉMATIQUES

La régression linéaire pour des données intervalles

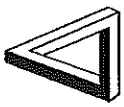
Vassart, Alice

Award date:
2006

FUNDP
Faculté des Sciences
Département de Mathématique

Rempart de la Vierge, 8
B-5000 Namur Belgique

La régression linéaire pour des données intervalles



Mémoire présenté pour l'obtention
du grade de
Licencié en Sciences Mathématiques
par

Alice VASSART

Promoteur : Prof. André HARDY

Année Académique 2005-2006

J'adresse tous mes remerciements à Monsieur André Hardy, professeur au Département de Mathématique aux FUNDP, pour l'aide précieuse et le soutien constructif qu'il m'a apportés dans l'élaboration de ce mémoire.

Abstract

The advent of computers makes possible today extremely large databases. We may then be interested in analysing classes of individuals called concepts instead of single individuals. Moreover, observations in a large data set can be studied more easily after aggregation to one of a smaller size.

The resulting observations will not be single-valued anymore, but interval-valued, multi-valued, histrograms or diagrams. These are called symbolic data.

In this paper, we extend classical linear regression to symbolic and more especially interval-valued data.

In the first part, the different types of symbolic data are introduced. We then study descriptive statistics for such data. These will be used to fit a symbolic linear regression model.

In the second part, we recall classical linear regression.

The third part concerns the fitting of a linear regression model to interval-valued data. We develop several methods : center method, lower bound and upper bound methods, center and range method and two other methods only for simple linear regression. We illustrate and compare these different methods by applying them on artificial and real data sets. We note that, among these methods, the center and range method and the center method seems to be the most efficient.

In the fourth part, we extend the center method to histogram-valued variables and we also propose a linear regression method in the case of explanatory diagram-valued variables.

In the fifth and last part, two applications are studied with the help of the module SREG of symbolic linear regression of the SODAS 2 software. We compare classical linear regression and linear regression at the level of concepts defined from the dependant variable. We notice that symbolic linear regression gives interesting results in comparison with linear regression on the first order individuals. In particular, Fisher and Student tests are less efficient in the presence of a large number of individuals in the regression.

However, we emphasize that these tests and the R square have not been rigorously extended to symbolic data.

Table des matières

Introduction

Les progrès de la technologie informatique permettent aujourd'hui le recueil de données en quantité très importante. Nous pouvons alors nous intéresser à l'étude de classes d'individus appelées concepts plutôt qu'à celle des individus du premier ordre. D'autre part, des masses considérables de données peuvent être plus facilement étudiées après réduction de leur taille à l'aide de concepts sous-jacents.

Ces concepts ne sont plus décrits par des valeurs uniques, mais par des ensembles de valeurs, des intervalles, des histogrammes ou des diagrammes, qui tiennent compte de la variation à l'intérieur des classes. De telles données sont appelées données symboliques.

L'objectif de ce mémoire est d'étendre la régression linéaire classique aux données symboliques et plus particulièrement de type intervalle.

Dans la première partie, nous introduisons les différents types de données symboliques. Nous présentons ensuite les statistiques descriptives pour de telles données, qui nous seront utiles pour étudier la régression linéaire symbolique.

La deuxième partie rappelle la régression linéaire classique.

La troisième partie est consacrée à l'extension de la régression linéaire au cas des variables intervalles. Nous commençons par présenter la méthode du centre de L. Billard et E. Diday programmée dans le logiciel SODAS 2. Cette méthode consiste à ajuster le modèle de régression linéaire classique sur le centre des intervalles. Nous reprenons aussi les méthodes de la borne inférieure et de la borne supérieure qui sont analogues à cette première méthode. Nous exposons ensuite la méthode du centre et de l'étendue proposée par F. A. T. De Carvalho, E. A. Lima Neto et C. P. Tenorio. Celle-ci consiste à ajuster deux modèles de régression linéaire classique, le premier permettant

d'estimer le centre d'un intervalle assumé par la variable dépendante, et le second son étendue.
Enfin, nous présentons d'autres méthodes proposées par O. Rodriguez uniquement pour la régression linéaire simple.

Nous illustrons et comparons ces différentes méthodes en les appliquant à des données artificielles et réelles.

Dans la quatrième partie, nous présentons l'extension de la méthode de L. Billard et E. Diday au cas des variables histogrammes, ainsi que la méthode de régression linéaire avec des variables explicatives diagrammes proposée par F. Afonso. Il s'agit des méthodes programmées dans le logiciel SODAS 2.

Nous terminons ce mémoire par une cinquième partie proposant deux applications traitées à l'aide du module SREG de SODAS 2. Nous comparons la régression linéaire classique et la régression linéaire au niveau de concepts définis à partir de la variable dépendante.

Première partie

Les données symboliques

Chapitre 1

Différents types de données symboliques

Les différents types de données symboliques sont décrits dans le chapitre 3 de [Bock00] dont nous présentons ici un résumé.

1.1 Variables multivaluée et intervalle

1.1.1 Variable multivaluée

Soit $E = \{1, ..., n\}$ un ensemble de n objets.

Une variable Y, dont l'espace d'observations est $\mathcal{Y}$, est dite **à valeurs dans un ensemble** $\mathcal{B}$ si

$$Y : E \to \mathcal{B} : k \rightsquigarrow Y(k)$$

où $\mathcal{B} = \mathcal{P}(\mathcal{Y}) = \{U \neq \emptyset | U \subseteq \mathcal{Y}\}$.

Une variable Y est **multivaluée** si ses valeurs $Y(k)$ sont toutes des sous-ensembles finis de $\mathcal{Y}$:

$$|Y(k)| < \infty \; \forall \; k \in E.$$

Une variable Y est **multivaluée catégorique** si $\mathcal{Y}$ a un nombre fini de catégories et donc

$$|Y(k)| < \infty \; \forall \; k \in E.$$

Une variable Y est **multivaluée quantitative** si ses valeurs $Y(k)$ sont des ensembles finis de nombres réels :

$$Y(k) \subseteq \mathbb{R} \text{ et } |Y(k)| < \infty \; \forall \; k \in E.$$

Exemple 1.1.1 *Variable multivaluée*

- E = ensemble des magasins d'une ville
- $Y_1(k)$: articles proposés par le magasin k
- $Y_2(k)$: chiffres d'affaires du magasin k en 2004 et en 2005.

On aura par exemple

$$Y_1(k) = \{\text{vêtements, chaussures, bijoux, parfums}\},$$

$$Y_2(k) = \{15000, 18000\}.$$

Dans cet exemple, Y_1 est une variable multivaluée *catégorique* et Y_2 est une variable multivaluée *quantitative.*

□

1.1.2 Variable de type intervalle

Une variable Y est de type **intervalle** si ses valeurs $Y(k)$ sont des intervalles de $\mathbb{R}$.

Dans ce cas, $\mathcal{B}$ est l'ensemble des intervalles de $\mathbb{R}$.

Exemple 1.1.2 *Variable de type intervalle*

- E = ensemble des habitants d'une ville
- Y : temps passé quotidiennement au téléphone (en minutes)

On peut avoir $Y(k) = [0, 10]$, $Y(l) = [10, 20]$,... où $k, l \in E$.

□

1.1.3 Variables multivaluée et intervalle par agrégation

Soit $\Omega = \{1, ..., n\}$ un ensemble d'individus appelés **objets du premier ordre**, sur lesquels on évalue une variable classique (univaluée) $\tilde{Y}$.

Considérons $E = \{C_1, ..., C_m\}$, un ensemble de classes $C_i \subseteq \Omega$ appelées **objets du second ordre.**
On se demande comment caractériser le comportement de ces classes par rapport à la variable $\tilde{Y}$.

Une solution est de définir une variable **agrégée** Y qui spécifie les valeurs prises par $\tilde{Y}$ sur les éléments de chaque classe C_i.

Exemple 1.1.3 *Variable multivaluée obtenue par agrégation*

- Ω = ensemble des étudiants inscrits en sciences mathématiques aux FUNDP
- $E = \{C_1, ..., C_5\}$ = ensemble des 5 années d'études
- $\tilde{Y}(k)$: taille de l'étudiant k en mètres

On peut décrire la première année de baccalauréat en mathématique par

$$Y(C_1) = \{1.50, 1.52, 1.55, 1.60, 1.63, 1.64, 1.66, 1.66, 1.71, 1.74, 1.74, 1.75, 1.79, 1.80\}.$$

□

Exemple 1.1.4 *Variable intervalle obtenue par agrégation*

Dans l'exemple précédent, on peut décrire l'année C_i par $Y(C_i) = [\alpha, \beta]$ où

$$\alpha = \min_{k \in C_i} \{\tilde{Y}(k)\},$$

$$\beta = \max_{k \in C_i} \{\tilde{Y}(k)\}.$$

On a donc pour la description de la première année de baccalauréat en mathématique

$$Y(C_1) = [1.50, 1.80].$$

□

1.2 Variable modale

Une variable **modale** Y sur un ensemble $E = \{1, ..., n\}$, dont l'espace d'observations est $\mathcal{Y}$, est définie par

$$Y(k) = (U(k), \pi_k) \quad \forall k \in E$$

où

- π_k est une mesure non-négative (une distribution de fréquence, de probabilité ou un poids) sur l'espace des observations $\mathcal{Y}$;
- $U(k) \subseteq \mathcal{Y}$ est le support de π_k dans le domaine $\mathcal{Y}$.

Exemple 1.2.1 *Variable modale*

- E = ensemble des pays européens
- $Y(k)$: langue(s) officielle(s) du pays k

Par exemple, si k désigne le pays "Belgique", on aura

$$Y(k) = \{(\text{néerlandais}, 0.50), (\text{français}, 0.45), (\text{allemand}, 0.05)\}.$$

Cela signifie que

- 50 % des Belges parlent le néerlandais,
- 45 % des Belges parlent le français,
- 5 % des Belges parlent l'allemand.

□

Chapitre 2

Statistiques descriptives pour des données symboliques

Après un bref rappel des statistiques descriptives classiques, nous étendons ces statistiques aux données symboliques. Nous considérons d'abord les statistiques unidimensionnelles et ensuite les statistiques à deux dimensions.

2.1 Statistiques descriptives unidimensionnelles

Pour une variable classique et pour des variables symboliques multivaluée et intervalle, nous reprenons l'approche de P. Bertrand et F. Goupil [Bertrand00].

L'étude d'une variable symbolique modale est issue des articles de L. Billard et E. Diday [Billard02b] et [Billard03].

2.1.1 Pour une variable classique

Soit $E = \{1, ..., n\}$ un ensemble de n individus décrits par une variable classique Y d'espace d'observations $\mathcal{Y}$.

Notons $y = (y_1, ..., y_n) = (Y(1), ..., Y(n))$ la séquence des valeurs prises par Y sur les individus $k \in E$.

Supposons que Y prenne l valeurs différentes sur E et notons ces valeurs $\xi_1, ..., \xi_l$.

Pour $i \in \{1, ..., l\}$, la **fréquence** de ξ_i est le nombre d'occurences de la valeur ξ_i :

$$n_i := \#\{k \in E | Y(k) = \xi_i\}.$$

La **distribution de fréquence** de Y est l'association des différentes valeurs prises par Y et de leurs fréquences correspondantes.

Pour $i \in \{1, ..., l\}$, la **fréquence relative** de ξ_i est définie par

$$f_i := \frac{n_i}{n}.$$

La **distribution statistique** de Y est l'association des différentes valeurs prises par Y sur E et de leurs fréquences relatives correspondantes.

On préfère représenter la distribution statistique de Y plutôt que sa distribution de fréquence (qui dépend du nombre d'individus n).

Si Y est une variable *qualitative* ou *quantitative discrète*, on représente habituellement sa distribution statistique par un **diagramme en bâtons**. Dans cette représentation graphique, chaque couple (ξ_i, f_i) est représenté par une bande verticale de base constante située en ξ_i le long de l'axe horizontal, et de hauteur proportionnelle à f_i.

Si Y est une variable *qualitative*, on peut également représenter sa distribution par un **diagramme circulaire** dans lequel chaque fréquence f_i est représentée par un secteur angulaire d'angle proportionnel à la fréquence.

Si Y est une variable *quantitative continue*, on regroupe les différentes valeurs prises par Y sur E en m intervalles consécutifs. Plus précisément, on ordonne les différentes valeurs ξ_i par ordre croissant

$$\xi_{(1)} \leq \xi_{(2)} \leq ... \leq \xi_{(l)}$$

et on considère une partition $\{I_1, ..., I_m\}$ de l'intervalle $\mathcal{I} = [\xi_{(1)}, \xi_{(l)}]$, où $I_j = [u_{j-1}, u_j[$ pour $j = 1, ..., m-1$ et $I_m = [u_{m-1}, u_m]$.

Notons

$$p_j = \frac{m_j}{n}$$

où m_j est le nombre de valeurs prises par Y dans l'intervalle I_j :

$$m_j = \#\{k \in E | Y(k) \in I_j\}.$$

On représente alors chaque couple (I_j, p_j) par un rectangle ayant pour base l'intervalle I_j le long de l'axe horizontal, et dont l'aire est proportionnelle à p_j. Une telle représentation graphique est appelée un **histogramme**.

Si Y est une variable *quantitative*, l'information donnée par sa distribution peut aussi être donnée par sa **fonction de répartition empirique** F_y définie par

$$F_y(\xi) := \frac{1}{n}\#\{k \in E | Y(k) \leq \xi\} \qquad \xi \in \mathbb{R}.$$

Pour les variables *quantitatives*, on utilise souvent la **moyenne arithmétique** qui est une mesure de centralité et l' **écart-type** qui est une mesure de dispersion.

Ces mesures sont définies respectivement par

$$\bar{y} := \frac{1}{n}\sum_{i=1}^{n} y_i$$

et

$$s_y := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Notons que la **variance**, définie par $s_y^2 := \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, peut encore s'écrire

$$s_y^2 = \frac{\sum_{i=1}^{n} y_i^2}{n} - \bar{y}^2.$$

La **fréquence observée** $\mathcal{O}_y$ de Y est définie par la fonction

$$\mathcal{O}_y : \mathcal{Y} \to \mathbb{N} : \xi \rightsquigarrow \mathcal{O}_y(\xi) := \#\{k \in E | Y(k) = \xi\}.$$

Par exemple, si Y dont l' espace d'observations $\mathcal{Y} = \mathbb{N}$ prend les valeurs $y = (6, 8, 9, 5, 8, 6)$ sur 6 individus, la fréquence observée de Y est

$$\mathcal{O}_y(\xi) = \begin{cases} 1 & si\, \xi \in \{5, 9\}, \\ 2 & si\, \xi \in \{6, 8\}, \\ 0 & sinon. \end{cases}$$

A partir du concept de fréquence observée, on peut facilement retrouver la fréquence n_i de la valeur ξ_i, ainsi que la fonction de répartition empirique F_y, la moyenne arithmétique $\bar{y}$ et l'écart-type s_y :

$n_i = \mathcal{O}_y(\xi_i) \qquad i = 1, ..., l\,;$

$F_y(\xi) = \frac{1}{n} \sum_{\xi_j \leq \xi} \mathcal{O}_y(\xi_j)\,;$

$\bar{y} = \frac{1}{n} \sum_{j=1}^{l} \mathcal{O}_y(\xi_j)\xi_j\,;$

$s_y = \sqrt{\frac{1}{n} \sum_{j=1}^{l} \mathcal{O}_y(\xi_j)(\xi_j - \bar{y})^2}\,.$

Nous allons maintenant définir un cadre qui permet l'extension symbolique de la définition classique de fréquence observée.

En généralisant les égalités ci-dessus, nous établirons alors les définitions de fonction de répartition empirique, ainsi que de moyenne arithmétique et d'écart-type dans le cas d'une variable symbolique.

2.1.2 Pour une variable symbolique

Le tableau des données symboliques

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits par p variables symboliques $Y_1, ..., Y_p$ d' espaces d'observations respectifs $\mathcal{Y}_1, ..., \mathcal{Y}_p$.

On note $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$.

Les données sont représentées par une matrice $\underline{X} = (\xi_{kj})$ de taille $(n \times p)$ où chaque cellule $\underline{X}(k,j)$ est la valeur prise par la variable Y_j sur l'objet $k \in E$:

$$\underline{X}(k,j) = \xi_{kj} = Y_j(k) \qquad k = 1,...,n;\ j = 1,...,p.$$

La ligne k de la matrice $\underline{X}$ nous donne donc la description symbolique de l'objet $k \in E$, c'est-à-dire le **vecteur de description** d_k défini par

$$d_k = (\underline{X}(k,1),...,\underline{X}(k,p))' = (\xi_{k1},...,\xi_{kp})' \qquad k = 1,...,n.$$

Considérons un vecteur de description arbitraire

$$d = (D_1,...,D_p)' \in \times_{j=1}^{p} \mathcal{P}(\mathcal{Y}_j)$$

où D_j est un élément quelconque de $\mathcal{P}(\mathcal{Y}_j)$ pour $j = 1,...,p$,

et notons $D = D_1 \times ... \times D_p \in \mathcal{P}(\mathcal{X})$ l'ensemble de description associé à d.

Un vecteur de description $d = (D_1,...,D_p)'$ est dit **individuel** si chaque $D_j \subseteq \mathcal{Y}_j$ est un singleton :

$$D_j = \{x_j\} \quad \text{avec } x_j \in \mathcal{Y}_j, \qquad j = 1,...,p.$$

Remarque

On identifiera souvent le vecteur de description individuel $d = (\{x_1\},...,\{x_p\})'$ et le vecteur $x = (x_1,...,x_p)' \in \mathcal{X}$. Plus généralement, on identifiera le vecteur de description $d = (D_1,...,D_p)'$ et son ensemble de description associé $D = D_1 \times ... \times D_p$.

Notation

Etant donné un vecteur de description individuel $x \in \mathcal{X}$, on notera ses composantes x_j par $x_{[Y_j]}$ pour $j = 1,...,p$. On a donc $x = (x_{[Y_1]},...,x_{[Y_p]})'$.

Dépendances logiques

Une dépendance logique peut être formulée au moyen d'une règle.

Par exemple, considérons les deux variables suivantes :

Y_1 : poids d'une personne en kg,
Y_2 : taille d'une personne en cm.

La règle définie par

$$\text{si [poids } \leq 55] \text{ alors [taille } \leq 180]$$

exprime une dépendance logique entre les valeurs de $\mathcal{X} = \mathcal{Y}_1 \times \mathcal{Y}_2$.

On exprime habituellement cette règle de la façon suivante :

$$[y_1 \leq 55] \Rightarrow [y_2 \leq 180]$$

où $y_1 \in \mathcal{Y}_1$ et $y_2 \in \mathcal{Y}_2$.

De façon générale, une dépendance logique sera exprimée au moyen d'une règle

$$v : [x \in A] \Rightarrow [x \in B]$$

où A et B sont deux ensembles de description de $\mathcal{X}$ et $x \in \mathcal{X}$ est un vecteur de description individuel.

En d'autres termes, un vecteur de description individuel x satisfait la règle v si et seulement si $x \in A \cap B$ ou $x \notin A$.

On écrira $v(x) = 1$ si x satisfait la règle v, et $v(x) = 0$ sinon.

Extension virtuelle d'un vecteur de description

On note $V_{\mathcal{X}}$ la collection de toutes les règles qui décrivent les dépendances logiques définies sur $\mathcal{X}$.

Soit $d = (D_1, ..., D_p)'$ un vecteur de description et soit $D = D_1 \times ... \times D_p$ son ensemble de description associé. L' **extension virtuelle** $vir(d)$ est la collection des vecteurs de description individuels de D qui satisfont toutes les règles de $V_{\mathcal{X}}$:

$$vir(d) = \{x \in D | v(x) = 1 \, \forall v \in V_{\mathcal{X}}\}.$$

Les éléments de $vir(d)$ sont appelés **vecteurs de description individuels réels.**

NB : Si $V_{\mathcal{X}} = \emptyset$, c'est-à-dire si aucune dépendance logique entre les variables n'a été définie, alors

$$vir(d) = \{x \in D_1 \times ... \times D_p\}$$

est l'ensemble des vecteurs de description individuels associés à la description d.

Exemple 2.1.1

Nous avons imaginé un tableau de données pour illustrer le concept d'extension virtuelle d'un vecteur de description.

Ce tableau contient les valeurs de deux variables multivaluées Y_1 et Y_2 sur les objets de l'ensemble $E = \{1, 2, 3, 4, 5\}$. Les espaces d'observations de Y_1 et Y_2 sont respectivement $\mathcal{Y}_1 = \{\text{rouge, bleu, noir}\}$ et $\mathcal{Y}_2 = \{1, 2\}$.

	Y_1	Y_2
1	{rouge, noir}	{2}
2	{bleu, noir}	{1}
3	{rouge, bleu}	{1, 2}
4	{rouge, bleu}	{2}
5	{bleu}	{1, 2}

Supposons qu'il existe une dépendance logique parmi les valeurs de $\mathcal{X} = \mathcal{Y}_1 \times \mathcal{Y}_2$, représentée par la règle

$$v : y_1 \in \{\text{rouge, bleu}\} \Rightarrow y_2 \in \{1\}.$$

Considérons la description symbolique de l'objet 1, c'est-à-dire le vecteur de description $d_1 = (\{\text{rouge, noir}\}, \{2\})'$.

L'extension virtuelle de d_1 est alors

$$vir(d_1) = \{x \in \{\text{rouge, noir}\} \times \{2\} | v(x) = 1\} = \{(\text{noir}, 2)'\}.$$

De la même manière, on trouve

$vir(d_2) = \{(\text{bleu}, 1)', (\text{noir}, 1)'\}$;
$vir(d_3) = \{(\text{rouge}, 1)', (\text{bleu}, 1)'\}$;
$vir(d_4) = \emptyset$;
$vir(d_5) = \{(\text{bleu}, 1)'\}$.

Remarquons que $vir(d_4) = \emptyset$ indique que l'objet 4 contredit la règle v. □

2.1.2.1 Cas d'une variable multivaluée

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits pas p variables multivaluées $Y_1, ..., Y_p$ catégoriques ou quantitatives.

On suppose que pour tout $k \in E$, le vecteur de description d_k a une extension virtuelle non vide :

$$\forall k \in E : |vir(d_k)| \neq 0.$$

En d'autres termes, aucun objet de E n'est en contradiction avec les règles de $V_{\mathcal{X}}$.

Considérons une des p variables multivaluées, notée Z, d'espace d'observations noté $\mathcal{Z}$.

La **fréquence observée** O_z de Z est la fonction définie par

$$O_z(\xi) := \sum_{k \in E} \pi_z(\xi; k) \qquad \xi \in \mathcal{Z}$$

où

$$\pi_z(\xi; k) := \frac{\#\{x \in vir(d_k) | x_{[Z]} = \xi\}}{|vir(d_k)|} \qquad \xi \in \mathcal{Z}, k \in E$$

est le pourcentage de vecteurs de description individuels $x \in vir(d_k)$ tels que $x_{[Z]} = \xi$.

On remarque que la fréquence observée est un *réel* positif et plus nécessairement un *entier* comme dans le cas classique.

Si Z est une variable classique (univaluée), le vecteur de description (individuel) d_k de tout $k \in E$ satisfait $d_{k[Z]} = Z(k)$. Donc, $|vir(d_k)| = 1$ si $vir(d_k)$ est non vide.

Par conséquent,

$$\pi_z(\xi; k) = \begin{cases} 1 & si\ d_{k[Z]} = Z(k) = \xi, \\ 0 & sinon. \end{cases}$$

On en déduit

$$\mathcal{O}_z(\xi) = \#\{k \in E | d_{k[Z]} = \xi\} = \#\{k \in E | Z(k) = \xi\},$$

ce qui coïncide avec la définition de fréquence observée que nous avons vue dans le cas classique (section 2.1.1).

◇

On peut facilement montrer que

$$\sum_{\xi \in \mathcal{Z}} \mathcal{O}_z(\xi) = n.$$

Remarque : définition alternative de fréquence observée

On peut inclure des poids pour chaque objet $k \in E$ dans la définition de fréquence observée :

$$\mathcal{O}_z(\xi) := \sum_{k \in E} w_k \pi_z(\xi; k) \qquad \xi \in \mathcal{Z}$$

avec $\sum_{k \in E} w_k = 1$ et $w_k \geq 0$ pour tout $k \in E$.

NB : Une telle définition de fréquence observée conduirait à des définitions de moyenne arithmétique et d'écart-type qui incluraient également des poids.

Par exemple, si on considère les vecteurs de description *individuels réels* comme unités élémentaires, la définition de $\mathcal{O}_z$ utilise les poids suivants :

$$w_k = \frac{|vir(d_k)|}{\sum_{k \in E} |vir(d_k)|} \qquad k \in E.$$

La définition alternative de fréquence observée est également recommandée dans le cas suivant. Supposons que les individus de Ω sont regroupés en *classes* $C_k \subseteq \Omega\,(k = 1, ..., n)$ considérées comme unités élémentaires décrites par p variables multivaluées.

Si on veut cependant considérer les *individus* de Ω comme unités élémentaires, on utilise la définition de $\mathcal{O}_z$ avec les poids suivants :

$$w_k = \frac{|C_k|}{|\Omega|} \qquad k = 1, ..., n.$$

La **distribution de fréquence observée** d'une variable multivaluée Z est l'association des différentes valeurs ξ de $\mathcal{Z}$ et de leurs fréquences observées $\mathcal{O}_z(\xi)$.

La **distribution statistique observée** de Z est l'association des différentes valeurs ξ de $\mathcal{Z}$ et des valeurs $\frac{\mathcal{O}_z(\xi)}{n}$ correspondantes.

Exemple 2.1.2

Considérons à nouveau le tableau de données de l'exemple 2.1.1.

Nous allons déterminer les distributions de fréquence et statistique observées de la variable Y_1 d'espace d'observations $\mathcal{Y}_1 = \{\text{rouge, bleu, noir}\}$.

Rappelons que nous avions obtenu

$vir(d_1) = \{(\text{noir}, 2)'\}$;
$vir(d_2) = \{(\text{bleu}, 1)', (\text{noir}, 1)'\}$;
$vir(d_3) = \{(\text{rouge}, 1)', (\text{bleu}, 1)'\}$;
$vir(d_4) = \emptyset$;
$vir(d_5) = \{(\text{bleu}, 1)'\}$.

Comme $vir(d_4) = \emptyset$, nous ne tenons pas compte de l'objet 4. Nous considérons donc l'ensemble $E' = E \setminus \{4\}$ avec $n' = |E'| = 4$.

On obtient alors

$$\mathcal{O}_{y_1}(\text{rouge}) = \sum_{k \in E'} \frac{\#\{x \in vir(d_k) | x_{[Y_1]} = \text{rouge}\}}{|vir(d_k)|} = \frac{0}{1} + \frac{0}{2} + \frac{1}{2} + \frac{0}{1} = 0.5.$$

De la même manière, on obtient

$\mathcal{O}_{y_1}(\text{bleu}) = 2$;
$\mathcal{O}_{y_1}(\text{noir}) = 1.5$.

La distribution de fréquence observée de Y_1 est donc

ξ	$\mathcal{O}_{y_1}(\xi)$
rouge	0.5
bleu	2
noir	1.5

Comme $n' = 4$, la distribution statistique observée de Y_1 est

ξ	$\frac{\mathcal{O}_{y_1}(\xi)}{n'}$
rouge	0.125
bleu	0.5
noir	0.375

□

Supposons maintenant que Z soit une variable multivaluée *quantitative* et notons $\xi_1, ..., \xi_l$ les éléments de l'ensemble $Q := \cup_{k \in E} Z(k) \subseteq \mathcal{Z} \subseteq \mathbb{R}$.

La **fonction de répartition empirique** de Z est définie par

$$F_z(\xi) := \frac{1}{n} \sum_{\xi_j \leq \xi} \mathcal{O}_z(\xi_j) \qquad \xi \in \mathbb{R}.$$

Sa **moyenne arithmétique** et son **écart-type** sont respectivement définis par

$$\bar{z} := \frac{1}{n} \sum_{j=1}^{l} \mathcal{O}_z(\xi_j)\xi_j$$

et

$$s_z := \sqrt{\frac{1}{n} \sum_{j=1}^{l} \mathcal{O}_z(\xi_j)(\xi_j - \bar{z})^2}.$$

Exemple 2.1.3

Reprenons à nouveau les données de l'exemple 2.1.1.

Nous allons déterminer la fonction de répartition empirique, ainsi que la moyenne et l'écart-type de la variable multivaluée *quantitative* Y_2.

A l'aide des extensions virtuelles des vecteurs de description, calculées précédemment, on obtient

$$\mathcal{O}_{y_2}(\xi) = \begin{cases} \frac{0}{1} + \frac{2}{2} + \frac{2}{2} + \frac{1}{1} = 3 & si\, \xi = 1, \\ \frac{1}{1} + \frac{0}{2} + \frac{0}{2} + \frac{0}{1} = 1 & si\, \xi = 2, \\ 0 & sinon. \end{cases}$$

Comme $n' = 4$, on obtient

$$F_{y_2}(\xi) = \begin{cases} 0 & si\, \xi < 1, \\ 0.75 & si\, 1 \leq \xi < 2, \\ 1 & si\, \xi \geq 2. \end{cases}$$

La moyenne arithmétique de Y_2 est donnée par

$$\bar{y}_2 = \frac{1}{4}\{3 \cdot 1 + 1 \cdot 2)\} = 1.25$$

et son écart-type est

$$s_{y_2} = \sqrt{\frac{1}{4}\{3(1 - 1.25)^2 + 1(2 - 1.25)^2\}} = \sqrt{0.1875} = 0.433.$$

□

2.1.2.2 Cas d'une variable intervalle

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits pas p variables symboliques $Y_1, ..., Y_p$.

Supposons qu'une de ces p variables symboliques, notée Z, soit une variable de type intervalle.

Pour chaque $k \in E$, on note $Z(k) = [\underline{z}_k, \bar{z}_k]$. Donc, $\underline{z}_k$ (respectivement $\bar{z}_k$) est la borne inférieure (respectivement supérieure) de l'intervalle $Z(k) \subseteq \mathbb{R}$.

Nous ne tenons pas compte du type des variables symboliques $Y_1, ..., Y_p$ différentes de Z, mais nous supposons qu'il n'existe pas de dépendances logiques entre les valeurs prises par Z et celles prises par les autres variables.

Notre approche est basée sur les hypothèses suivantes :

i) Chaque objet $k \in E$ a été sélectionné avec la même probabilité $\frac{1}{n}$.

ii) Pour tout $k \in E$, les valeurs $x_{[Z]}$ de tout vecteur de description individuel $x \in vir(d_k)$ sont uniformément distribuées dans l'intervalle $Z(k) = [\underline{z}_k, \bar{z}_k]$.

Autrement dit, $x_{[Z]}$ (avec $x \in \cup_{k \in E} vir(d_k)$), version univaluée de Z, est considérée comme une variable aléatoire dont la distribution est le mélange uniforme de n distributions $U[\underline{z}_k, \bar{z}_k]$, $k = 1, ..., n$.

Si $x \in vir(d_k)$, la fonction de répartition de $x_{[Z]}$ est donc la fonction de répartition d'une loi $U[\underline{z}_k, \bar{z}_k]$:

$\forall \quad \xi \in \mathbb{R}$:

$$P(x_{[Z]} \leq \xi | x \in vir(d_k)) = \begin{cases} 0 & si\, \xi < \underline{z}_k, \\ \frac{\xi - \underline{z}_k}{\bar{z}_k - \underline{z}_k} & si\, \underline{z}_k \leq \xi < \bar{z}_k, \\ 1 & si\, \bar{z}_k \leq \xi. \end{cases}$$

La **fonction de répartition empirique** de Z est la fonction de répartition du mélange uniforme de n distributions $U[\underline{z}_k, \bar{z}_k]$, $k = 1, ..., n$:

$$F_z(\xi) := \frac{1}{n} \sum_{k \in E} P(x_{[Z]} \leq \xi | x \in vir(d_k)) \qquad \xi \in \mathbb{R}.$$

En remplaçant $P(x_{[Z]} \leq \xi | x \in vir(d_k))$ par son expression donnée ci-dessus, on obtient

$$F_z(\xi) = \frac{1}{n} \sum_{k : \xi \in Z(k)} \frac{\xi - \underline{z}_k}{\bar{z}_k - \underline{z}_k} + \frac{1}{n} \#\{k \in E | \xi \geq \bar{z}_k\} \qquad \xi \in \mathbb{R}.$$

En dérivant la fonction de répartition empirique de Z par rapport à ξ, on obtient sa **fonction de densité empirique** :

$$f_z(\xi) := \frac{1}{n} \sum_{k:\xi \in Z(k)} \frac{1}{\bar{z}_k - \underline{z}_k} = \frac{1}{n} \sum_{k \in E} \frac{\mathbb{1}_{Z(k)}(\xi)}{l(Z(k))} \qquad \xi \in \mathbb{R}$$

où $\mathbb{1}_A$ désigne la fonction indicatrice de l'ensemble A et $l(I)$ est la longueur de l'intervalle I.

NB : La fonction de densité empirique de Z, donnée par f_z pour un échantillon E de n objets satisfaisant les hypothèses i) et ii), est seulement une approximation de la limite pour $n \to \infty$ de la distribution de $x_{[Z]}^{(n)}$, version univaluée de Z définie sur un échantillon de taille n.

Pour construire un histogramme de Z, notons

$$\mathcal{I} = [min\{\underline{z}_k | k \in E\}, max\{\bar{z}_k | k \in E\}]$$

et considérons une partition de $\mathcal{I}$ en m intervalles $I_j = [u_{j-1}, u_j[$ pour $j = 1, ..., m-1$ et $I_m = [u_{m-1}, u_m]$.

Soit p_j la probabilité que $x_{[Z]}$ appartienne à I_j, où x est un vecteur de description individuel quelconque.

Sous nos hypothèses, on a

$$p_j = P(x_{[Z]} \in I_j) = \frac{1}{n} \sum_{k \in E} \frac{l(Z(k) \cap I_j)}{l(Z(k))}.$$

L' **histogramme** de Z associé à la partition de $\mathcal{I}$ en $\{I_1, ..., I_m\}$ est la représentation de chaque couple (I_j, p_j) par un rectangle ayant pour base l'intervalle I_j le long de l'axe horizontal, et dont l'aire est proportionnelle à p_j.

La **moyenne** de Z est définie par

$$\bar{z} := \int_{-\infty}^{+\infty} \xi f_z(\xi)\, d\xi\,.$$

Son **écart-type** est

$$s_z := \sqrt{\int_{-\infty}^{+\infty} (\xi - \bar{z})^2 f_z(\xi)\, d\xi}.$$

Plus précisément, en remplaçant $f_z(\xi)$ par son expression, on obtient

$$\bar{z} = \frac{1}{n} \sum_{k \in E} \int_{-\infty}^{+\infty} \frac{\mathbb{1}_{Z(k)}(\xi)}{l(Z(k))} \xi \, d\xi$$

$$= \frac{1}{n} \sum_{k \in E} \int_{\underline{z}_k}^{\bar{z}_k} \frac{\xi}{l(Z(k))} \, d\xi$$

$$= \frac{1}{n} \sum_{k \in E} \frac{1}{(\bar{z}_k - \underline{z}_k)} \left[\frac{\xi^2}{2} \right]_{\underline{z}_k}^{\bar{z}_k}$$

$$= \frac{1}{n} \sum_{k \in E} \frac{\bar{z}_k^2 - \underline{z}_k^2}{2\,(\bar{z}_k - \underline{z}_k)}.$$

On a donc

$$\bar{z} = \frac{1}{2n} \sum_{k \in E} (\bar{z}_k + \underline{z}_k).$$

D'autre part, l'écart-type de Z peut encore s'écrire

$$s_z = \sqrt{\int_{-\infty}^{+\infty} (\xi^2 + \bar{z}^2 - 2\xi\bar{z}) f_z(\xi) \, d\xi}$$

$$= \sqrt{\int_{-\infty}^{+\infty} \xi^2 f_z(\xi) \, d\xi + \bar{z}^2 - 2\bar{z}\bar{z}}$$

$$= \sqrt{\int_{-\infty}^{+\infty} \xi^2 f_z(\xi) \, d\xi - \bar{z}^2}.$$

En remplaçant $f_z(\xi)$ par son expression, l'intégrale (moment d'ordre 2) s'écrit

$$\int_{-\infty}^{+\infty} \xi^2 f_z(\xi)\, d\xi = \frac{1}{n} \sum_{k \in E} \int_{\underline{z}_k}^{\bar{z}_k} \frac{\xi^2}{l(Z(k))}\, d\xi$$

$$= \frac{1}{n} \sum_{k \in E} \frac{1}{(\bar{z}_k - \underline{z}_k)} \left[\frac{\xi^3}{3} \right]_{\underline{z}_k}^{\bar{z}_k}$$

$$= \frac{1}{n} \sum_{k \in E} \frac{\bar{z}_k^3 - \underline{z}_k^3}{3\,(\bar{z}_k - \underline{z}_k)}$$

$$= \frac{1}{3n} \sum_{k \in E} (\bar{z}_k^2 + \bar{z}_k \underline{z}_k + \underline{z}_k^2).$$

On obtient alors

$$s_z = \sqrt{\frac{1}{3n} \sum_{k \in E} (\bar{z}_k^2 + \bar{z}_k \underline{z}_k + \underline{z}_k^2) - \frac{1}{4n^2} \left[\sum_{k \in E} (\bar{z}_k + \underline{z}_k) \right]^2}.$$

Exemple 2.1.4

Nous développons l'exemple mentionné dans [Billard02b] par L. Billard et E. Diday et concernant le tableau de données suivant issu de [Raju97].

Ce tableau contient les valeurs (intervalles) de la fréquence cardiaque (nombre de battements cardiaques par minute) Y de 10 patients :

k	Y
1	[44, 68]
2	[60, 72]
3	[56, 90]
4	[70, 112]
5	[54, 72]
6	[70, 100]
7	[72, 100]
8	[76, 98]
9	[86, 96]
10	[86, 100]

Afin de construire un histogramme de Y, déterminons

$$min\{\underline{y}_k | k \in E\} = 44 \text{ et } max\{\bar{y}_k | k \in E\} = 112\,.$$

Prenons $\mathcal{I} = [40, 120]$ et considérons la partition de $\mathcal{I}$ en les 8 intervalles consécutifs $I_1 = [40, 50[, I_2 = [50, 60[, ..., I_7 = [100, 110[, I_8 = [110, 120]$.

Considérons l'intervalle $I_4 = [70, 80[$ et déterminons la probabilité p_4 que $x_{[Y]}$ appartienne à I_4, où x est un vecteur de description individuel quelconque.

On a

$$p_4 = \frac{1}{10}\left(0 + \frac{2}{12} + \frac{10}{34} + \frac{10}{42} + \frac{2}{18} + \frac{10}{30} + \frac{8}{28} + \frac{4}{22} + 0 + 0\right) = 0.1611.$$

Le couple (I_4, p_4) sera représenté sur l'histogramme par un rectangle ayant pour base l'intervalle I_4 le long de l'axe horizontal, et dont l'aire est proportionnelle à p_4.

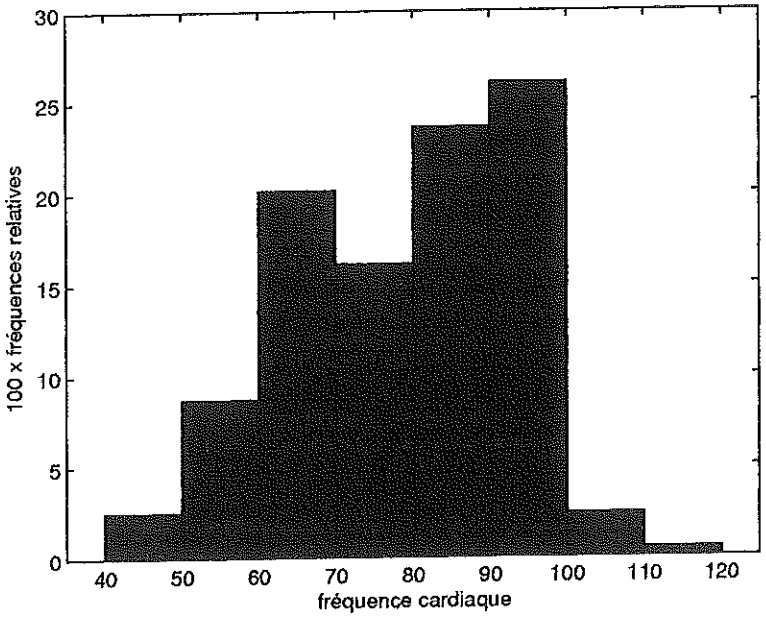
NB : Comme tous les intervalles ont la même longueur, la hauteur de chaque rectangle représentant un couple (I_j, p_j) est égale à la fréquence relative p_j.

De la même manière, nous avons calculé les probabilités p_j pour tous les autres intervalles I_j :

$$p_1 = 0.0250; \qquad p_2 = 0.0868; \qquad p_3 = 0.2016;$$

$$p_5 = 0.2363; \qquad p_6 = 0.2606; \qquad p_7 = 0.0238; \qquad p_8 = 0.0048.$$

Nous avons alors représenté l'histogramme de Y :



Nous avons ensuite calculé la moyenne de Y :

$$\begin{aligned}\bar{y} &= \frac{1}{2 \cdot 10}(112 + 132 + 146 + 182 + 126 + 170 + 172 + 174 + 182 + 186) \\ &= 79.1\end{aligned}$$

et son écart-type :

$$s_y = \sqrt{\tfrac{1}{3 \cdot 10} 194180 - 79.1^2} = \sqrt{215.857} = 14.692.$$

□

2.1.2.3 Cas d'une variable modale multi-catégorique (ou diagramme)

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits pas p variables modales multi-catégoriques $Y_1, ..., Y_p$ d'espaces d'observations respectifs $\mathcal{Y}_1, ..., \mathcal{Y}_p$.

Supposons qu'il existe une dépendance logique exprimée par la règle v entre les valeurs de $\mathcal{X} = \times_{j=1}^{p} \mathcal{Y}_j$.

On suppose que pour tout $k \in E$, le vecteur de description d_k a une extension virtuelle non vide :

$$\forall k \in E : |vir(d_k)| \neq 0.$$

En d'autres termes, aucun objet de E n'est en contradiction avec la règle v.

Considérons une des p variables modales multi-catégoriques que nous notons Z : pour tout $k \in E$, $Z(k)$ prend les valeurs ξ_j avec, respectivement, les probabilités p_{kj}, $j = 1, ..., s_k$, où $\sum_{j=1}^{s_k} p_{kj} = 1$.

NB : $Z(k)$ peut donc être représentée par un diagramme pour tout objet $k \in E$.

La **fréquence observée** $\mathcal{O}_z$ de Z est la fonction définie par

$$\mathcal{O}_z(\xi_j) := \sum_{k \in E} \pi_z(\xi_j; k) \qquad j = 1, ..., s_k$$

où

$$\pi_z(\xi_j; k) := P(Z = \xi_j | v(x) = 1, k)$$

$$= \frac{\sum_x P(x_{[Z]} = \xi_j | v(x) = 1, k)}{\sum_x \sum_{j=1}^{s_k} P(x_{[Z]} = \xi_j | v(x) = 1, k)}$$

où, pour tout $k \in E$, $P(x_{[Z]} = \xi_j | v(x) = 1, k)$ est la probabilité que $x_{[Z]} = \xi_j$ lorsque la règle v est satisfaite et x est un vecteur de description individuel associé à la description symbolique d_k de l'objet k.

On peut montrer que

$$\sum_{j=1}^{s_k} \mathcal{O}_z(\xi_j) = n.$$

Exemple 2.1.5

Considérons les variables classiques qualitatives suivantes :

$\tilde{Y}_1$: type de chauffage central, $\mathcal{Y}_1 = \{$gaz, combustible solide, électricité, autre$\}$,
$\tilde{Y}_2$: chauffage central installé, $\mathcal{Y}_2 = \{$non, oui$\}$.

Dans le cadre d'un recensement du "U.K. Office for National Statistics", ces deux variables ont été mesurées sur les maisons de 374 parties du Royaume-Uni. Par agrégation, on obtient les données du tableau suivant qui représente les valeurs (distributions de probabilité) des variables agrégées Y_1 et Y_2 dans 25 régions du pays.

	Y_1				Y_2	
k	gaz	combustible solide	électricité	autre	non	oui
1	{.87	.07.	.05	.01}	{.09	.91}
2	{.71	.11	.10	.08}	{.12	.88}
3	{.83	.08	.09	.00}	{.23	.77}
4	{.76	.06	.11	.07}	{.19	.81}
5	{.78	.06	.09	.07}	{.12	.88}
6	{.90	.01	.08	.01}	{.22	.78}
7	{.87	.01	.10	.02}	{.22	.78}
8	{.78	.02	.13	.07}	{.13	.87}
9	{.91	.00	.09	.00}	{.24	.76}
10	{.73	.08	.11	.08}	{.14	.86}
11	{.59	.07	.17	.17}	{.10	.90}
12	{.90	.01	.08	.01}	{.19	.71}
13	{.84	.00	.14	.02}	{.09	.91}
14	{.82	.00	.11	.07}	{.17	.83}
15	{.88	.00	.09	.03}	{.12	.88}
16	{.85	.01	.10	.04}	{.09	.91}
17	{.71	.03	.17	.09}	{.16	.83}
18	{.87	.09	.04	.00}	{.13	.87}
19	{.32	.24	.24	.20}	{.25	.75}
20	{.50	.12	.28	.10}	{14	.86}
21	{.69	.13	.18	.00}	{.12	.88}
22	{.79	.01	.20	.00}	{.21	.79}
23	{.72	.05	.19	.04}	{.07	.93}
24	{.43	.00	.43	.14}	{.28	.72}
25	{.00	.41	.14	.45}	{.09	.91}

On suppose que les valeurs $Y_1(k)$ $(k = 1, ..., 25)$ représentent les types de chauffage central des maisons de la région k *dans lesquelles le chauffage central est installé.*

Par exemple, la région 1 est telle que 87% des maisons dans lesquelles le chauffage central est installé se chauffent au gaz, 7% utilisent un combustible solide, 5% l'électricité et 1% un autre type de chauffage central.
De plus, 9% des maisons de la région 1 n'ont pas le chauffage central, tandis que 91% l'ont.

Supposons dans un premier temps qu'il n'y a pas de règle logique v à satisfaire.

Nous nous intéressons tout d'abord à la variable Y_1.

Notons ξ_1 = gaz, ξ_2 = combustible solide, ξ_3 = électricité et ξ_4 = autre.

Calculons la fréquence observée de Y_1 en ξ_1 :

$$\mathcal{O}_{y_1}(\xi_1) = 0.87 + 0.71 + ... + 0.43 + 0.00 = 18.05.$$

De la même manière, on obtient

$$\mathcal{O}_{y_1}(\xi_2) = 1.67, \quad \mathcal{O}_{y_1}(\xi_3) = 3.51, \quad \mathcal{O}_{y_1}(\xi_4) = 1.77.$$

La distribution statistique de Y_1 est donc :

ξ_j	$\frac{\mathcal{O}_{y_1}(\xi_j)}{n}$
ξ_1	0.722
ξ_2	0.067
ξ_2	0.140
ξ_2	0.071

De façon similaire, on obtient la distribution statisique de Y_2 :

ξ	$\frac{\mathcal{O}_{y_2}(\xi)}{n}$
non	0.156
oui	0.844

Supposons maintenant que la règle logique suivante doit être satisfaite :

$$v = \begin{cases} v_1 : y_2 \in \{non\} \Rightarrow y_1 \in \{\xi_0\}, \\ v_2 : y_2 \in \{oui\} \Rightarrow y_1 \in \{\xi_1, ..., \xi_4\}, \end{cases}$$

où on a noté $\tilde{Y}_1 = \xi_0$ si aucun type de chauffage central n'est utilisé.

Autrement dit, une maison qui a le chauffage central doit utiliser un des types de chauffage ξ_j, $j = 1, ..., 4$.

Les vecteurs de description individuels qui satisfont la règle v appartiennent à $\{(\xi_0, non), (\xi_1, oui), (\xi_2, oui), (\xi_3, oui), (\xi_4, oui)\}$.

Calculons dans ce cas la fréquence observée de Y_1 en ξ_1 :

$$\mathcal{O}_{y_1}(\xi_1) = 0.7917 + 0.6248 + ... + 0.3096 + 0 = 15.225.$$

Nous avons aussi calculé

$$\mathcal{O}_{y_1}(\xi_0) = 3.9, \quad \mathcal{O}_{y_1}(\xi_2) = 1.45, \quad \mathcal{O}_{y_1}(\xi_3) = 2.925, \quad \mathcal{O}_{y_1}(\xi_4) = 1.5.$$

La distribution statistique de Y_1 est alors :

ξ_j	$\frac{\mathcal{O}_{y_1}(\xi_j)}{n}$
ξ_0	0.156
ξ_1	0.609
ξ_2	0.058
ξ_3	0.117
ξ_4	0.060

La distribution statistique de Y_2 est :

ξ	$\frac{\mathcal{O}_{y_2}(\xi)}{n}$
non	0.156
oui	0.844

Remarque : L. Billard et E. Diday étudient cet exemple dans [Billard02b] en remplaçant la description symbolique de l'objet $u = 25$ par

$$\{\{0.25\,\xi_0, 0.00\,\xi_1, 0.41\,\xi_2, 0.14\,\xi_3, 0.20\,\xi_4\}, \{0.09, 0.91\}\}\,.$$

□

2.1.2.4 Cas d'une variable modale intervalle (ou histogramme)

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits par p variables symboliques $Y_1, ..., Y_p$.

Supposons qu'une de ces variables, notée Z, soit une variable modale intervalle : pour chaque $k \in E$, $Z(k)$ prend les valeurs $[\underline{z}_{kj}, \bar{z}_{kj}[$ avec, respectivement, les probabilités p_{kj}, $j = 1, ..., s_k$, où $\sum_{j=1}^{s_k} p_{kj} = 1$.

NB : $Z(k)$ est donc un histogramme pour tout objet $k \in E$.

Comme dans le cas d'une variable intervalle, on suppose qu'il n'y a pas de dépendances logiques entre les valeurs prises par Z et celles prises par les autres variables, et que les hypothèses suivantes sont vérifiées :

i) Chaque objet $k \in E$ a été selectionné avec la même probabilité $\frac{1}{n}$.

ii) Pour tout $k \in E$ et pour tout intervalle $[\underline{z}_{kj}, \bar{z}_{kj}[$, les valeurs $x_{[Z]}$ de tout vecteur de description individuel $x \in vir(d_k)$ sont uniformément distribuées dans cet intervalle.

On a donc, pour tout $\xi \in \mathbb{R}$:

$$P(x_{[Z]} \leq \xi | x \in vir(d_k)) = \begin{cases} 0 & si\ \xi < \underline{z}_{kj}, \\ \frac{\xi - \underline{z}_{kj}}{\bar{z}_{kj} - \underline{z}_{kj}} & si\ \underline{z}_{kj} \leq \xi < \bar{z}_{kj}, \\ 1 & si\ \bar{z}_{kj} \leq \xi. \end{cases}$$

Pour construire un histogramme de Z, déterminons

$$\mathcal{I} = \left[min\left\{\underline{z}_{kj} | k \in E,\, j \in \{1, ..., s_k\}\right\}, max\left\{\bar{z}_{kj} | k \in E,\, j \in \{1, ..., s_k\}\right\}\right]$$

et considérons une partition de $\mathcal{I}$ en m intervalles $I_i = [u_{i-1}, u_i[$ pour $i = 1, ..., m-1$ et $I_m = [u_{m-1}, u_m]$.

La fréquence observée de Z sur l'intervalle I_i est

$$\mathcal{O}_z(i) = \sum_{k \in E} \pi_z(i;k)$$

où

$$\pi_z(i;k) = \sum_{j \in Z(i)} \frac{l(Z(k;j) \cap I_i)}{l(Z(k;j))} p_{kj}$$

où $Z(k;j)$ est l'intervalle $[\underline{z}_{kj}, \bar{z}_{kj}[$ et $Z(i)$ est l'ensemble de tous les intervalles $Z(k;j)$ qui ont une intersection non vide avec I_i, pour un k donné.

On a

$$\sum_{i=1}^{m} \mathcal{O}_z(i) = n.$$

La fréquence relative de l'intervalle I_i est

$$p_i = \frac{\mathcal{O}_z(i)}{n}.$$

L'**histogramme** de Z est la représentation de chaque couple (I_i, p_i) par un rectangle ayant pour base l'intervalle I_i le long de l'axe horizontal, et dont l'aire est proportionnelle à p_i.

La **fonction de densité empirique** de Z est définie par

$$f_z(\xi) := \frac{1}{n} \sum_{k \in E} \sum_{j=1}^{s_k} \frac{\mathbb{1}_{Z(k;j)}(\xi)}{l(Z(k;j))} p_{kj} \qquad \xi \in \mathbb{R}.$$

La **moyenne de Z** est

$$\bar{z} := \frac{1}{2n} \sum_{k \in E} \sum_{j=1}^{s_k} (\underline{z}_{kj} + \bar{z}_{kj}) p_{kj}$$

et son **écart-type** est

$$s_z := \sqrt{\frac{1}{3n}\sum_{k\in E}\sum_{j=1}^{s_k}(\bar{z}_{kj}^2 + \bar{z}_{kj}\underline{z}_{kj} + \underline{z}_{kj}^2)p_{kj} - \frac{1}{4n^2}\left[\sum_{k\in E}\sum_{j=1}^{s_k}(\bar{z}_{kj} + \underline{z}_{kj})p_{kj}\right]^2}.$$

Exemple 2.1.6

Considérons la variable classique continue $\tilde{Y}(k)$: poids de la femme k en kg.

Supposons que le poids $\tilde{Y}$ d'un ensemble de femmes ait été mesuré et que ces femmes aient été regroupées selon les 7 tranches d'âges $[20,30[$, $[30,40[$, $[40,50[$, $[50,60[$, $[60,70[$, $[70,80[$ et $[80,90[$ notées respectivement $1,2,3,4,5,6$ et 7.

Voici les valeurs (histogrammes) de la variable agrégée Y sur ces 7 classes :

$Y(1) = \{[35,42[,.02;\ [42,48[,.06;\ [48,54[,.24;\ [54,60[,.30;\ [60,66[,.24;\ [66,72[,.06;\ [72,80[,.08\}$;

$Y(2) = \{[50,54[,.02;\ [54,58[,.06;\ [58,62[,.40;\ [62,66[,.24;\ [66,70[,.24;\ [70,75[,.04\}$;

$Y(3) = \{[55,62[,.04;\ [62,67[,.14;\ [67,72[,.20;\ [72,77[,42;\ [77,82[,.14;\ [82,89[,.04;\ [89,94[,.02\}$;

$Y(4) = \{[50,57[,.04;\ [57,63[,.06;\ [63,69[,.20;\ [69,75[,.26;\ [75,81[,.28;\ [81,87[,.12;\ [87,95[,.04\}$;

$Y(5) = \{[63,68[,.04;\ [68,72[,.14;\ [72,76[,.38;\ [76,80[,.22;\ [80,84[,.16;\ [84,90[,.06\}$;

$Y(6) = \{[67,72[,.04;\ [72,75[,.06;\ [75,78[,.24;\ [78,81[,.26;\ [81,84[,.22;\ [84,87[,.14;\ [87,91[.04\}$;

$Y(7) = \{[50,60[,.02;\ [60,67[,.12;\ [67,74[,.16;\ [74,81[,.24;\ [81,88[,.32;\ [88,95[,.10;\ [95,102[.04\}$.

Par exemple, 40% des femmes qui ont entre 30 et 40 ans pèsent entre 58 et 62 kg.

Remarque : Les données de cet exemple ont été construites en s'inspirant de celles traitées par L. Billard et E. Diday dans [Billard02b].

Pour construire un histogramme de Y, déterminons

$$min\{\underline{y}_{kj}|k \in E, j \in \{1,...,s_k\}\} = 35$$

et

$$max\{\bar{y}_{kj}|k \in E, j \in \{1,...,s_k\}\} = 102.$$

Prenons $\mathcal{I} = [35, 105]$ et considérons la partition de $\mathcal{I}$ en les 10 intervalles consécutifs $I_1 = [35, 42[$, $I_2 = [42, 49[$, $I_3 = [49, 56[$,...,$I_9 = [91, 98[$, $I_{10} = [98, 105]$.

Considérons par exemple le troisième intervalle $I_3 = [49, 56[$.

On a

$\pi_y(3;1) = \frac{5}{6}\,0.24 + \frac{2}{6}\,0.30 = 0.3$;

$\pi_y(3;2) = \frac{4}{4}\,0.02 + \frac{2}{4}\,0.06 = 0.05$;

$\pi_y(3;3) = \frac{1}{7}\,0.04 = 0.0057$;

$\pi_y(3;4) = \frac{6}{7}\,0.04 = 0.0343$;

$\pi_y(3;5) = 0$;

$\pi_y(3;6) = 0$;

$\pi_y(3;7) = \frac{6}{10}\,0.02 = 0.012$.

On obtient alors la fréquence observée de l'intervalle I_3 :

$$\mathcal{O}_y(3) = \sum_{k \in E} \pi_y(3;k) = 0.402$$

et sa fréquence relative :

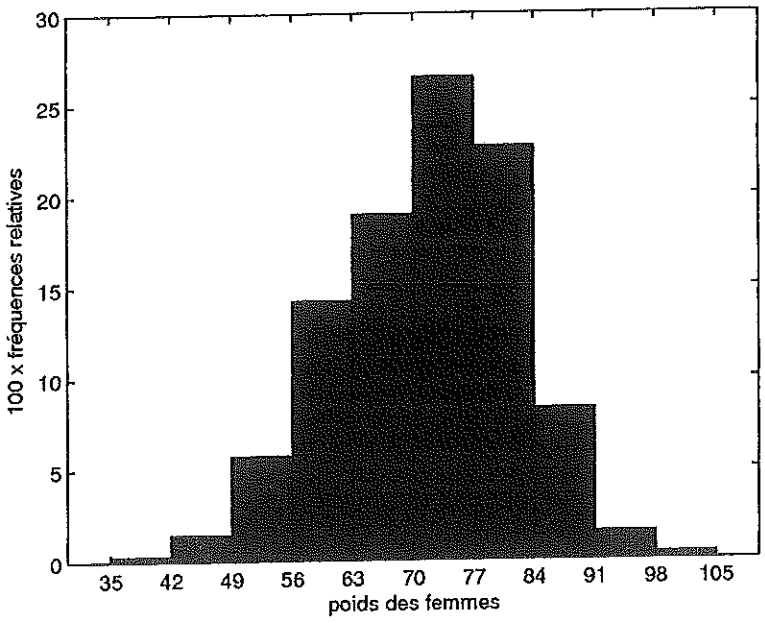
$$p_3 = \frac{\mathcal{O}_y(3)}{n} = \frac{0.402}{7} = 0.05743.$$

Le couple (I_3, p_3) sera représenté sur l'histogramme par un rectangle ayant pour base l'intervalle I_3 le long de l'axe horizontal, et dont l'aire est proportionnelle à p_3.

Nous avons repris les fréquences observées et relatives des tous les intervalles I_i $(i = 1, ..., 10)$ dans le tableau suivant :

i	I_i	$\mathcal{O}_y(i)$	$p_i = \frac{\mathcal{O}_y(i)}{n}$
1	$[35, 42[$	0.02	0.00286
2	$[42, 49[$	0.1	0.01429
3	$[49, 56[$	0.402	0.05743
4	$[56, 63[$	0.997	0.14243
5	$[63, 70[$	1.326	0.18943
6	$[70, 77[$	1.855	0.265
7	$[77, 84[$	1.587	0.22671
8	$[84, 91[$	0.582	0.08314
9	$[91, 98[$	0.106	0.01514
10	$[98, 105]$	0.023	0.00329

Nous avons alors représenté l'histogramme de Y :



Nous avons ensuite calculé la moyenne de Y :

$$\bar{y} = \frac{1}{2 \cdot 7} 1003.36 = 71.669$$

et sont écart-type :

$$s_y = \sqrt{\frac{1}{3 \cdot 7} 110149.56 - 71.669^2} = 10.429.$$

□

2.2 Statistiques descriptives à deux dimensions

2.2.1 Pour deux variables classiques

Cette section est basée sur le cours de "Statistiques" de A. Hardy [Hardy03].

Soit $E = \{1, ..., n\}$ un ensemble de n individus décrits par 2 variables classiques X et Y d'espaces d'observations respectifs $\mathcal{X}$ et $\mathcal{Y}$.

Supposons que X prenne k valeurs différentes sur E notées $x_1, ..., x_k$, et que Y prenne l valeurs différentes sur E qu'on note $y_1, ..., y_l$.

La **fréquence conjointe** de x_i et y_j est définie par

$$n_{ij} := \#\{k \in E | X(k) = x_i \text{ et } Y(k) = y_j\} \qquad i = 1, ..., k;\ j = 1, ..., l.$$

La **distribution de fréquence conjointe** de X et Y est l'association des différents couples (x_i, y_j) $(i = 1, ..., k; j = 1, ..., l)$ et de leurs fréquences conjointes correspondantes.

La **fréquence relative conjointe** de x_i et y_j est

$$f_{ij} := \frac{n_{ij}}{n} \qquad i = 1, ..., k;\ j = 1, ..., l.$$

La **distribution statistique conjointe** de X et Y est l'association des différents couples (x_i, y_j) $(i = 1, ..., k; j = 1, ..., l)$ et de leurs fréquences relatives conjointes correspondantes.

On représente habituellement la distribution statistique conjointe de X et Y par un graphique à trois dimensions dans lequel l'axe vertical mesure les fréquences relatives conjointes, les deux autres axes horizontaux représentant les variables X et Y.

Si X et Y sont des variables *quantitatives*, on peut mesurer l'intensité de l'association linéaire entre X et Y à l'aide du **coefficient de corrélation linéaire de Pearson** défini par

$$r_{xy} := \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} = \frac{s_{xy}}{s_x s_y}$$

où

$$s_{xy} := \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

est la **covariance** entre X et Y,

$\bar{x}$ et $\bar{y}$ sont les moyennes arithmétiques de X et de Y, et s_x et s_y, leurs écarts-types.

Notons que la covariance entre X et Y peut encore s'écrire

$$s_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}.$$

La **fréquence conjointe observée** $\mathcal{O}_{x,y}$ de X et Y est la fonction définie par

$\mathcal{O}_{x,y} : \mathcal{X}\times\mathcal{Y} \to \mathbb{N} : (x_i, y_j) \to \mathcal{O}_{x,y}(x_i, y_j) := \#\{k \in E | X(k) = x_i \text{ et } Y(k) = y_j\}$.

On a

$$n_{ij} = \mathcal{O}_{x,y}(x_i, y_j) \qquad i = 1, ..., k;\ j = 1, ..., l;$$

$$s_{xy} = \left[\frac{1}{n}\sum_{i,j}(x_i \cdot y_j)\mathcal{O}_{x,y}(x_i, y_j)\right] - \bar{x}\bar{y}.$$

2.2.2 Pour deux variables symboliques

Cette section constitue un résumé des articles de L. Billard et E. Diday [Billard02b] et [Billard03].

2.2.2.1 Cas de deux variables multivaluées

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits pas p variables multivaluées $Y_1, ..., Y_p$.

On suppose que pour tout $k \in E$, le vecteur de description d_k a une extension virtuelle non vide :

$$\forall k \in E : |vir(d_k)| \neq 0.$$

Considérons deux des p variables multivaluées, notées Z_1 et Z_2, d'espaces d'observations respectifs $\mathcal{Z}_1$ et $\mathcal{Z}_2$.

La **fréquence conjointe observée** $\mathcal{O}_{z_1,z_2}$ de Z_1 et Z_2 est définie par

$$\mathcal{O}_{z_1,z_2}(\xi_1, \xi_2) := \sum_{k \in E} \pi_{z_1,z_2}(\xi_1, \xi_2; k) \qquad \xi_1 \in \mathcal{Z}_1 \text{ et } \xi_2 \in \mathcal{Z}_2$$

où

$$\pi_{z_1,z_2}(\xi_1, \xi_2; k) := \frac{\#\{x \in vir(d_k) | x_{[Z_1]} = \xi_1, x_{[Z_2]} = \xi_2\}}{|vir(d_k)|}$$

est le pourcentage de vecteurs de description individuels $x \in vir(d_k)$ tels que $x_{[Z_1]} = \xi_1$ et $x_{[Z_2]} = \xi_2$.

On peut montrer que

$$\sum_{\xi_1 \in \mathcal{Z}_1, \xi_2 \in \mathcal{Z}_2} \mathcal{O}_{z_1,z_2}(\xi_1, \xi_2) = n.$$

La **distribution de fréquence conjointe observée** de Z_1 et Z_2 est l'association des différents couples $(\xi_1, \xi_2) \in \mathcal{Z}_1 \times \mathcal{Z}_2$ et de leurs fréquences conjointes observées $\mathcal{O}_{z_1,z_2}(\xi_1, \xi_2)$.

La **distribution statistique conjointe observée** de Z_1 et Z_2 est l'association des différents couples $(\xi_1, \xi_2) \in \mathcal{Z}_1 \times \mathcal{Z}_2$ et des valeurs $\frac{\mathcal{O}_{z_1,z_2}(\xi_1,\xi_2)}{n}$ correspondantes.

Si Z_1 et Z_2 sont des variables multivaluées *quantitatives*, on peut définir la covariance et le coefficient de corrélation entre Z_1 et Z_2.

La **covariance** entre Z_1 et Z_2 est définie par

$$s_{z_1 z_2} := \left[\frac{1}{n} \sum_{\xi_1,\xi_2} (\xi_1 \cdot \xi_2) \mathcal{O}_{z_1,z_2}(\xi_1, \xi_2) \right] - \bar{z}_1 \bar{z}_2$$

où $\bar{z}_1$ et $\bar{z}_2$ sont les moyennes des variables multivaluées Z_1 et Z_2, définies dans la section 2.1.3.

Le **coefficient de corrélation** entre Z_1 et Z_2 est alors

$$r_{z_1 z_2} := \frac{s_{z_1 z_2}}{s_{z_1} s_{z_2}}$$

où s_{z_1} et s_{z_2} sont les écarts-types des variables multivaluées Z_1 et Z_2, définis dans la section 2.1.3.

Exemple 2.2.1

Nous reprenons les données utilisées par L. Billard et E. Diday dans [Billard02b].

Considérons les variables classiques suivantes :

$\tilde{Y}_1$: présence de cancer, $\mathcal{Y}_1 = \{\text{non} = 0, \text{oui} = 1\}$,

$\tilde{Y}_2$: nombre de traitements relatifs au cancer, $\mathcal{Y}_2 = \{0, 1, 2, 3\}$.

Supposons que ces variables ont été mesurées sur les patients d'un hôpital qui ont été agrégés en 9 classes de telle sorte qu'on obtienne le tableau de données suivant :

k	Y_1	Y_2
1	$\{0,1\}$	$\{2\}$
2	$\{0,1\}$	$\{0,1\}$
3	$\{0,1\}$	$\{3\}$
4	$\{0,1\}$	$\{2,3\}$
5	$\{0\}$	$\{1\}$
6	$\{0\}$	$\{0,1\}$
7	$\{1\}$	$\{2,3\}$
8	$\{1\}$	$\{1,2\}$
9	$\{1\}$	$\{1,3\}$

Supposons qu'il y ait une dépendance logique entre les valeurs de $\mathcal{Y}_1 \times \mathcal{Y}_2$, formulée par la règle

$$v : y_1 \in \{0\} \Rightarrow y_2 \in \{0\}.$$

Autrement dit, si aucun patient d'une classe de patients n'a eu de cancer, alors aucun patient de cette classe de patients n'a subi de traitement relatif au cancer.

Calculons les extensions virtuelles des vecteurs de description des 9 classes du tableau de données. On a

$vir(d_1) = \{x \in \{0,1\} \times \{(1,2)\} | v(x) = 1\} = \{(1,2)\}$;
$vir(d_2) = \{(0,0),(1,0),(1,1)\}$;
$vir(d_3) = \{(1,3)\}$;
$vir(d_4) = \{(1,2),(1,3)\}$;
$vir(d_5) = \emptyset$;
$vir(d_6) = \{(0,0)\}$;
$vir(d_7) = \{(1,2),(1,3)\}$;
$vir(d_8) = \{(1,1),(1,2)\}$;
$vir(d_9) = \{(1,1),(1,3)\}$.

Comme $vir(d_5) = \emptyset$, nous ne tenons pas compte de l'individu 5. On considère donc l'ensemble $E' = E \setminus \{5\}$ avec $n' = |E'| = 8$.

On calcule alors la distribution de fréquence conjointe observée de Y_1 et Y_2.

On a

$$\mathcal{O}_{y_1,y_2}(0,0) = \frac{0}{1} + \frac{1}{3} + \frac{0}{1} + \frac{0}{2} + \frac{1}{1} + \frac{0}{2} + \frac{0}{2} + \frac{0}{2} = \frac{4}{3};$$

$$\mathcal{O}_{y_1,y_2}(0,1) = 0;\ \mathcal{O}_{y_1,y_2}(0,2) = 0;\ \mathcal{O}_{y_1,y_2}(0,3) = 0;$$

$$\mathcal{O}_{y_1,y_2}(1,0) = \frac{1}{3};\ \mathcal{O}_{y_1,y_2}(1,1) = \frac{4}{3};\ \mathcal{O}_{y_1,y_2}(1,2) = \frac{5}{2};\ \mathcal{O}_{y_1,y_2}(1,3) = \frac{5}{2}.$$

Les moyennes et les écarts-types de Y_1 et Y_2 sont

$$\bar{y}_1 = \frac{5}{6}; \qquad \bar{y}_2 = \frac{83}{48};$$

$$s_{y_1} = 0.373; \qquad s_{y_2} = 1.113.$$

On obtient donc la covariance entre Y_1 et Y_2 :

$$s_{y_1y_2} = \frac{1}{8}\left[\left(1 \cdot \frac{4}{3}\right) + \left(2 \cdot \frac{5}{2}\right) + \left(3 \cdot \frac{5}{2}\right)\right] - \frac{5}{6} \cdot \frac{83}{48} = 0.288$$

et le coefficient de corrélation entre Y_1 et Y_2 :

$$r_{y_1y_2} = \frac{0.288}{0.373 \cdot 1.113} = 0.694.$$

□

2.2.2.2 Cas de deux variables intervalles

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits par p variables symboliques $Y_1, ..., Y_p$.

Supposons que deux de ces variables, notées Z_1 et Z_2, soient des variables de type intervalle.
Pour tout $k \in E$, on note $Z_i(k) = [\underline{z}_{ik}, \bar{z}_{ik}]$, $i = 1, 2$.

Pour chaque variable Z_i $(i = 1, 2)$, nous faisons les mêmes hypothèses que dans l'étude unidimensionnelle d'une variable intervalle (section 2.1.4).

La **fonction de densité empirique conjointe** de Z_1 et Z_2 est définie par

$$f_{z_1,z_2}(\xi_1, \xi_2) := \frac{1}{n} \sum_{k \in E} \frac{\mathbb{1}_{Z'(k)}(\xi_1, \xi_2)}{A(Z'(k))} \qquad \xi_1, \xi_2 \in \mathbb{R}$$

où $Z'(k)$ est le rectangle $Z_1(k) \times Z_2(k) = [\underline{z}_{1k}, \bar{z}_{1k}] \times [\underline{z}_{2k}, \bar{z}_{2k}]$ et $A(Z'(k))$ est l'aire de ce rectangle.

Pour construire l'histogramme conjoint de Z_1 et Z_2, on partitionne

$$\mathcal{I}_i = [min\{\underline{z}_{ik} | k \in E\}, max\{\bar{z}_{ik} | k \in E\}]$$

en m_i intervalles $I_{j_i} = [u_{j_i - 1}, u_{j_i}[$ pour $j_i = 1, ..., m_i - 1$ et $I_{m_i} = [u_{m_i - 1}, u_{m_i}]$, $i = 1, 2$.

Notons

$$R_{j_1 j_2} = I_{j_1} \times I_{j_2} \qquad j_1 = 1, ..., m_1; j_2 = 1, ..., m_2.$$

Soit $p_{j_1 j_2}$ la probabilité que pour un vecteur de description individuel quelconque x, $(x_{[Z_1]}, x_{[Z_2]}) \in R_{j_1 j_2}$.

On a

$$p_{j_1 j_2} = \frac{1}{n} \sum_{k \in E} \frac{A(Z'(k) \cap R_{j_1 j_2})}{A(Z'(k))}.$$

L' **histogramme conjoint** de Z_1 et Z_2 est la représentation de chaque couple $(R_{j_1 j_2}, p_{j_1 j_2})$ par un parallélépipède rectangle dont la base est le rectangle $R_{j_1 j_2}$, et dont le volume est proportionnel à $p_{j_1 j_2}$.

La **covariance** entre Z_1 et Z_2 est

$$
\begin{aligned}
s_{z_1 z_2} &:= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\xi_1 - \bar{z}_1)(\xi_2 - \bar{z}_2) f_{z_1,z_2}(\xi_1, \xi_2)\, d\xi_1 d\xi_2 \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (\xi_1\xi_2 - \xi_1\bar{z}_2 - \xi_2\bar{z}_1 + \bar{z}_1\bar{z}_2)\, f_{z_1,z_2}(\xi_1, \xi_2)\, d\xi_1 d\xi_2 \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \xi_1\xi_2 f_{z_1,z_2}(\xi_1, \xi_2)\, d\xi_1 d\xi_2 - \bar{z}_1\bar{z}_2 - \bar{z}_2\bar{z}_1 + \bar{z}_1\bar{z}_2 \\
&= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \xi_1\xi_2 f_{z_1,z_2}(\xi_1, \xi_2)\, d\xi_1 d\xi_2 - \bar{z}_1\bar{z}_2.
\end{aligned}
$$

En remplaçant $f_{z_1,z_2}(\xi_1, \xi_2)$ par son expression donnée ci-dessus, on obtient

$$
\begin{aligned}
s_{z_1 z_2} &= \frac{1}{n} \sum_{k \in E} \frac{1}{(\bar{z}_{1k} - \underline{z}_{1k})(\bar{z}_{2k} - \underline{z}_{2k})} \int_{\underline{z}_{1k}}^{\bar{z}_{1k}} \int_{\underline{z}_{2k}}^{\bar{z}_{2k}} \xi_1\xi_2\, d\xi_1 d\xi_2 - \bar{z}_1\bar{z}_2 \\
&= \frac{1}{n} \sum_{k \in E} \frac{1}{(\bar{z}_{1k} - \underline{z}_{1k})(\bar{z}_{2k} - \underline{z}_{2k})} \int_{\underline{z}_{1k}}^{\bar{z}_{1k}} \xi_1\, d\xi_1 \int_{\underline{z}_{2k}}^{\bar{z}_{2k}} \xi_2\, d\xi_2 - \bar{z}_1\bar{z}_2 \\
&= \frac{1}{n} \sum_{k \in E} \frac{1}{(\bar{z}_{1k} - \underline{z}_{1k})(\bar{z}_{2k} - \underline{z}_{2k})} \left[\frac{\xi_1^2}{2}\right]_{\underline{z}_{1k}}^{\bar{z}_{1k}} \left[\frac{\xi_2^2}{2}\right]_{\underline{z}_{2k}}^{\bar{z}_{2k}} - \bar{z}_1\bar{z}_2 \\
&= \frac{1}{4n} \sum_{k \in E} \frac{(\bar{z}_{1k}^2 - \underline{z}_{1k}^2)(\bar{z}_{2k}^2 - \underline{z}_{2k}^2)}{(\bar{z}_{1k} - \underline{z}_{1k})(\bar{z}_{2k} - \underline{z}_{2k})} - \bar{z}_1\bar{z}_2.
\end{aligned}
$$

On a donc

$$s_{z_1z_2} = \frac{1}{4n}\sum_{k\in E}(\bar{z}_{1k}+\underline{z}_{1k})(\bar{z}_{2k}+\underline{z}_{2k}) - \frac{1}{4n^2}\left[\sum_{k\in E}(\bar{z}_{1k}+\underline{z}_{1k})\right]\left[\sum_{k\in E}(\bar{z}_{2k}+\underline{z}_{2k})\right]$$

où les moyennes des variables intervalles Z_i $(i = 1, 2)$:

$$\bar{z}_i = \frac{1}{2n}\sum_{k\in E}(\bar{z}_{ik}+\underline{z}_{ik})$$

ont été définies dans la section 2.1.4.

Le **coefficient de corrélation** entre Z_1 et Z_2 est alors

$$r_{z_1z_2} := \frac{s_{z_1z_2}}{s_{z_1}s_{z_2}}$$

où s_{z_1} et s_{z_2} sont les écarts-types des variables intervalles Z_1 et Z_2, définis dans la section 2.1.4.

Exemple 2.2.2

Nous reprenons les données étudiées par L. Billard et E. Diday dans [Billard02b] et issues de [Raju97].

Considérons les variables

Y_1 : pression artérielle systolique (en mm Hg),

Y_2 : pression artérielle diastolique (en mm Hg).

Le tableau suivant contient les valeurs (intervalles) des pressions artérielles systolique Y_1 et diastolique Y_2 de 10 patients :

k	Y_1	Y_2
1	[90, 110]	[50, 70]
2	[90, 130]	[70, 90]
3	[140, 180]	[90, 100]
4	[110, 142]	[80, 108]
5	[90, 100]	[50, 70]
6	[134, 142]	[80, 110]
7	[130, 160]	[76, 90]
8	[110, 190]	[70, 110]
9	[138, 180]	[90, 110]
10	[110, 150]	[78, 100]

Nous allons construire l'histogramme conjoint de Y_1 et Y_2.

On partitionne

$$\mathcal{I}_1 = [min\{\underline{z}_{1k} | k \in E\}, max\{\bar{z}_{1k} | k \in E\}] = [90, 190]$$

en 10 intervalles consécutifs $I_1 = [90, 100[$, $I_2 = [100, 110[$, ... , $I_9 = [170, 180[$ et $I_{10} = [180, 190]$.

On partitionne également

$$\mathcal{I}_2 = [min\{\underline{z}_{2k} | k \in E\}, max\{\bar{z}_{2k} | k \in E\}] = [50, 110]$$

en 6 intervalles consécutifs $I_1 = [50, 60[$, $I_2 = [60, 70[$, ... , $I_5 = [90, 100[$ et $I_6 = [100, 110]$.

Calculons ensuite les probabilités $p_{j_1 j_2}$ que pour un vecteur de description individuel quelconque x, $(x_{[Y_1]}, x_{[Y_2]})$ appartienne à $R_{j_1 j_2} = [u_{j_1 - 1}, u_{j_1}[\times [u_{j_2 - 1}, u_{j_2}[$, $j_1 = 1, ..., 10$, $j_2 = 1, ..., 6$.

Considérons par exemple $R_{64} = [140, 150[\times [80, 90[$.

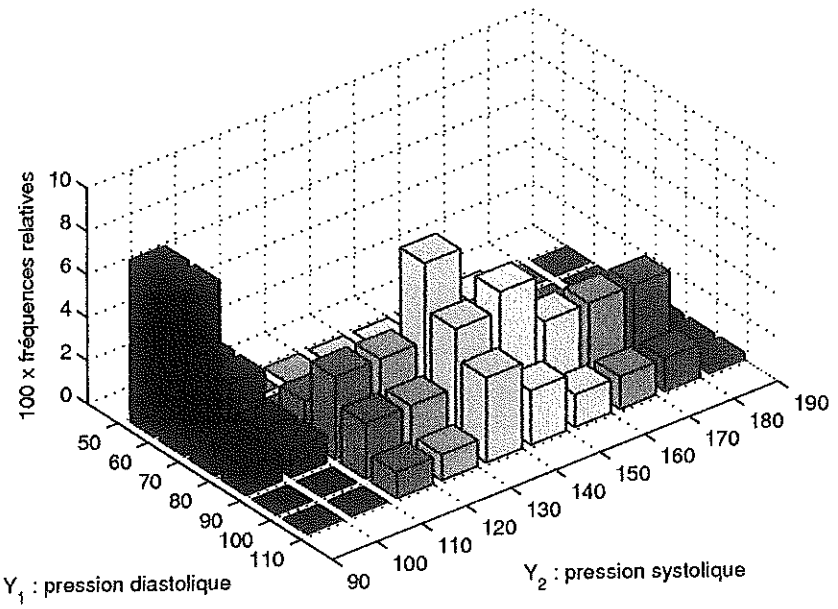
On a

$$p_{64} = \frac{1}{10}\left[0+0+0+\frac{2\cdot 10}{32\cdot 28}+0+\frac{2\cdot 10}{8\cdot 30}+\frac{10\cdot 10}{30\cdot 14}+\frac{10\cdot 10}{80\cdot 40}+0+\frac{10\cdot 10}{40\cdot 22}\right]$$
$$= 0.04886.$$

Nous avons repris les probabilités $p_{j_1 j_2} \times 10^2$ $(j_1 = 1, ..., 10; j_2 = 1, ..., 6)$ dans le tableau ci-dessous :

	j_2	1	2	3	4	5	6
j_1	$I_{j_1} \setminus I_{j_2}$	$[50,60[$	$[60,70[$	$[70,80[$	$[80,90[$	$[90,100[$	$[100,110]$
1	$[90,100[$	7.5	7.5	1.25	1.25	0	0
2	$[100,110[$	2.5	2.5	1.25	1.25	0	0
3	$[110,120[$	0	0	1.79	3.815	2.565	1.205
4	$[120,130[$	0	0	1.79	3.815	2.565	1.205
5	$[130,140[$	0	0	1.492	7.446	5.303	3.943
6	$[140,150[$	0	0	1.492	4.886	6.196	2.515
7	$[150,160[$	0	0	1.265	2.693	4.003	1.503
8	$[160,170[$	0	0	0.313	0.313	4.003	1.503
9	$[170,180[$	0	0	0.313	0.313	4.003	1.503
10	$[180,190]$	0	0	0.313	0.313	0.313	0.313

Nous avons alors représenté l'histogramme conjoint de Y_1 et Y_2 :



Les moyennes et les écarts-types de Y_1 et Y_2 sont

$$\bar{y}_1 = 131.3; \qquad \bar{y}_2 = 84.6;$$

$$s_{y_1} = 24.98; \qquad s_{y_2} = 15.15.$$

On obtient alors la covariance entre Y_1 et Y_2 :

$$s_{y_1 y_2} = \frac{1}{4 \cdot 10} 454636 - 131.3 \cdot 84.6 = 257.92$$

et le coefficient de corrélation entre Y_1 et Y_2 :

$$r_{y_1 y_2} = \frac{257.92}{24.98 \cdot 15.15} = 0.68.$$

□

2.2.2.3 Cas de deux variables modales intervalles (ou histogrammes)

Soit $E = \{1, ..., n\}$ un ensemble de n objets décrits par p variables symboliques $Y_1, ..., Y_p$.

Supposons que deux de ces variables, notées Z_i $(i = 1, 2)$, soient des variables modales intervalles.

Pour chaque $k \in E$, $Z_i(k)$ $(i = 1, 2)$ prend les valeurs $[\underline{z}_{ikj}, \bar{z}_{ikj}[$ avec, respectivement, les probabilités p_{ikj}, $j = 1, ..., s_{ik}$, où $\sum_{j=1}^{s_{ik}} p_{ikj} = 1$.

NB : $Z_i(k)$ est donc un histogramme pour chaque $k \in E$, $i = 1, 2$.

Pour chaque variable Z_i $(i = 1, 2)$, nous faisons les mêmes hypothèses que dans l'étude unidimensionnelle d'une variable histogramme (section 2.1.6).

La **fonction de densité empirique conjointe** de Z_1 et Z_2 est définie par

$$f_{z_1,z_2}(\xi_1,\xi_2) := \frac{1}{n}\sum_{k\in E}\left\{\sum_{j_1=1}^{s_{1k}}\sum_{j_2=1}^{s_{2k}}\frac{\mathbb{1}_{Z(k;j_1,j_2)}(\xi_1,\xi_2)}{A(Z(k;j_1,j_2))}p_{1kj_1}p_{2kj_2}\right\}$$

où $Z(k;j_1,j_2)$ est le rectangle $[\underline{z}_{1kj_1},\bar{z}_{1kj_1}[\times[\underline{z}_{2kj_2},\bar{z}_{2kj_2}[$ et $A(Z(k;j_1,j_2))$ est l'aire de ce rectangle.

Pour construire l'histogramme conjoint de Z_1 et Z_2, on partitionne

$$\mathcal{I}_i = \left[min\left\{\underline{z}_{ikj}|k\in E, j\in\{1,...,s_{ik}\}\right\}, max\left\{\bar{z}_{ikj}|k\in E, j\in\{1,...,s_{ik}\}\right\}\right]$$

en m_i intervalles $I_{l_i} = [u_{l_i-1}, u_{l_1}[$ pour $l_i = 1,...,m_i-1$ et $I_{m_i} = [u_{m_i-1}, u_{m_i}]$, $i = 1,2$.

Notons

$$R_{l_1l_2} = I_{l_1}\times I_{l_2} \qquad l_1 = 1,...,m_1; l_2 = 1,...,m_2.$$

Soit $p_{l_1l_2}$ la probabilité que pour un vecteur de description individuel quelconque x, $\left(x_{[Z_1]}, x_{[Z_2]}\right) \in R_{l_1l_2}$.

On a

$$p_{l_1l_2} = \frac{1}{n}\sum_{k\in E}\sum_{j_1\in Z(l_1)}\sum_{j_2\in Z(l_2)}\frac{A(Z(k;j_1,j_2)\cap R_{l_1l_2})}{A(Z(k;j_1,j_2))}p_{1kj_1}p_{2kj_2}$$

où $Z(l_i)$ représente l'ensemble des intervalles $Z(k;j_i) = [\underline{z}_{ikj_i}, \bar{z}_{ikj_i}[$, $i = 1,2$, qui ont une intersection non vide avec I_{l_i}, pour un k donné.

L' **histogramme conjoint** de Z_1 et Z_2 est la représentation de chaque couple $(R_{l_1l_2}, p_{l_1l_2})$ par un parallélépipède rectangle dont la base est le rectangle $R_{l_1l_2}$, et dont le volume est proportionnel à $p_{l_1l_2}$.

La **covariance** entre Z_1 et Z_2 est définie par

$$s_{z_1 z_2} := \frac{1}{4n} \sum_{k \in E} \left\{ \sum_{j_1=1}^{s_{1k}} \sum_{j_2=1}^{s_{2k}} (\bar{z}_{1kj_1} + \underline{z}_{1kj_1})(\bar{z}_{2kj_2} + \underline{z}_{2kj_2}) p_{1kj_1} p_{2kj_2} \right\}$$
$$- \frac{1}{4n^2} \left[\sum_{k \in E} \left\{ \sum_{j_1=1}^{s_{1k}} (\bar{z}_{1kj_1} + \underline{z}_{1kj_1}) p_{1kj_1} \right\} \right] \left[\sum_{k \in E} \left\{ \sum_{j_2=1}^{s_{2k}} (\bar{z}_{2kj_2} + \underline{z}_{2kj_2}) p_{2kj_2} \right\} \right]$$

où les moyennes des variables histogrammes Z_i $(i = 1, 2)$:

$$\bar{z}_i = \frac{1}{2n} \sum_{k \in E} \sum_{j_i=1}^{s_{ik}} (\bar{z}_{ikj_i} + \underline{z}_{ikj_i}) p_{ikj_i}$$

ont été définies dans la section 2.1.6.

Le **coefficient de corrélation** entre Z_1 et Z_2 est alors

$$r_{z_1 z_2} := \frac{s_{z_1 z_2}}{s_{z_1} s_{z_2}}$$

où s_{z_1} et s_{z_2} sont les écarts-types des variables histogrammes Z_1 et Z_2, définis dans la section 2.1.6.

Exemple 2.2.3

Considérons les variables

Y_1 : hématocrite (pourcentage du volume du sang occupé par des cellules rouges),

Y_2 : quantité totale d'hémoglobine dans le sang en grammes / décilitre.

Le tableau suivant contient les valeurs (histogrammes) de l'hématocrite Y_1 et de la quantité d'hémoglobine dans le sang Y_2 de 5 classes de patients :

k	Y_1 : hématocrite	Y_2 : hémoglobine
1	$\{[33.2, 37.8[, .7 ; [37.8, 39.6], .3\}$	$\{[11.5, 12.1[, .4 ; [12.1, 12.8], .6\}$
2	$\{[36.3, 38.1[, .2 ; [38.1, 47.4], .8\}$	$\{[12.3, 14.2[, .5 ; [14.2, 14.9[, .4 ; [14.9, 15.2], .1\}$
3	$\{[41.5, 44.9[, .5 ; [44.9, 48.8], .5\}$	$\{[14.3, 14.8[, .5 ; [14.8, 15.5[, .4 ; 15.5, .1\}$
4	$\{[44.4, 47.1[, .4 ; [47.1, 52.5], .6\}$	$\{[15.3, 15.7[, .3 ; [15.7, 16.4[, .5 ; [16.4, 16.7], .2\}$
5	$\{[39.3, 51.5], 1\}$	$\{[13.7, 14.5[, .5 ; [14.5, 15.2[, .3 ; [15.2, 16.5], .2\}$

L'histogramme conjoint de ces 2 variables histogrammes est représenté dans l'article de L. Billard et E. Diday [Billard03]. Nous en exposons ici la construction.

Partitionnons $\mathcal{I}_1 = [30, 55]$ en les 5 intervalles consécutifs $[30, 35[$, $[35, 40[$, ..., $[50, 55]$, et partitionnons $I_2 = [11, 17]$ en les 6 intervalles consécutifs $[11, 12[$, $[12, 13[$, ... , $[16, 17]$.

Calculons ensuite les probabilités $p_{l_1 l_2}$ que pour un vecteur de description individuel quelconque x, $(x_{[Y_1]}, x_{[Y_2]})$ appartienne à $R_{l_1 l_2} = [u_{l_1-1}, u_{l_1}[\times[u_{l_2-1}, u_{l_2}[$, $l_1 = 1, ..., 5$, $l_2 = 1, ..., 6$.

Par exemple, considérons $R_{21} = [35, 40[\times[11, 12[$.

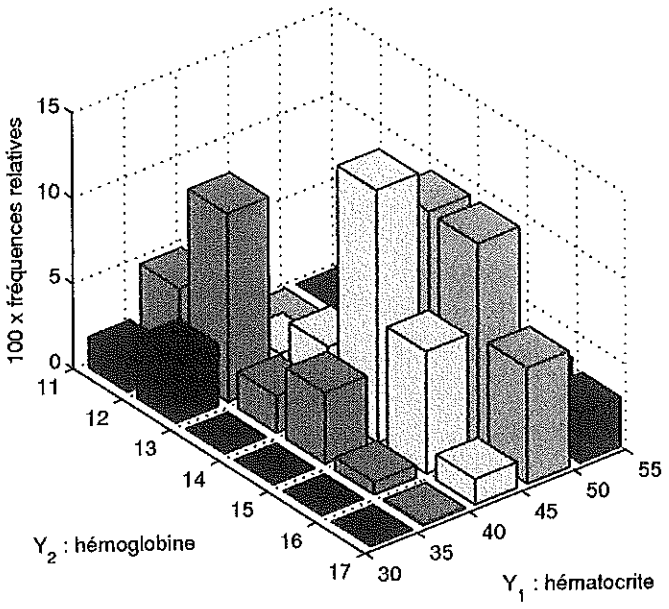
On a

$$p_{21} = \frac{1}{5}\left\{\left[\frac{2 \cdot 0.5}{4.6 \cdot 0.6} 0.7 \cdot 0.4 + 0 + \frac{1.8 \cdot 0.5}{1.8 \cdot 0.6} 0.3 \cdot 0.4 + 0\right] + 0 + 0 + 0 + 0\right\}$$
$$= 0.04841.$$

Nous avons repris les probabilités $p_{l_1 l_2} \times 10^2$ $(l_1 = 1, ..., 5; l_2 = 1, ..., 6)$ dans le tableau ci-dessous :

	l_2	1	2	3	4	5	6
l_1	$I_{l_1} \setminus I_{l_2}$	$[11, 12[$	$[12, 13[$	$[13, 14[$	$[14, 15[$	$[15, 16[$	$[16, 17]$
1	$[30, 35[$	1.826	3.652	0	0	0	0
2	$[35, 40[$	4.841	11.020	2.128	4.139	0.724	0.088
3	$[40, 45[$	0	1.585	3.801	14.799	7.155	1.494
4	$[45, 50[$	0	0.761	2.623	12.310	12.259	6.783
5	$[50, 55]$	0	0	0.461	1.295	3.371	2.888

Nous avons alors représenté l'histogramme conjoint de Y_1 et Y_2 :



Nous avons aussi calculé les moyennes et les écarts-types de Y_1 et de Y_2 :

$$\bar{y}_1 = 43.341\,; \quad \bar{y}_2 = 14.337\,;$$

$$s_{y_1} = 4.846\,; \quad s_{y_2} = 1.381\,.$$

La covariance et le coefficient de corrélation entre Y_1 et Y_2 sont donc

$$s_{y_1 y_2} = 5.005\,;$$

$$r_{y_1 y_2} = 0.748.$$

□

Deuxième partie

La régression linéaire pour des données classiques

Chapitre 3

Méthodologie

Nous présentons ici un résumé du cours de "Modèles Statistiques Linéaires" de A. Hardy [Hardy04].

3.1 Le modèle et ses hypothèses

On souhaite expliquer une variable aléatoire Y à partir de $p-1$ variables $X_1, ..., X_{p-1}$ par une relation linéaire

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \mathcal{E}$$

où $\mathcal{E}$ représente le terme d'erreur.

Y est dite **variable à expliquer** ou **variable dépendante** et $X_1, ..., X_{p-1}$ sont appelées **variables explicatives, variables indépendantes** ou encore **régresseurs.**

Soit $E = \{1, ..., n\}$ un ensemble de n individus sur lesquels on a mesuré ces p variables. On a donc n relations linéaires

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad i = 1, ..., n$$

où $y_i = Y(i)$ et $x_{ij} = X_j(i)$.

Ces n équations peuvent s'écrire sous forme matricielle

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ 1 & x_{21} & \dots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_n \end{pmatrix}$$

ou encore

$$Y = X\beta + \mathcal{E}$$

où

- Y est un vecteur aléatoire observable de taille n,
- X est une matrice connue de taille $(n \times p)$ appelée **matrice de régression,**
- β est un vecteur de paramètres inconnus de taille p,
- $\mathcal{E}$ est un vecteur aléatoire inobservable de taille n.

On suppose que les hypothèses suivantes sont vérifiées :

i) *Le rang de X est égal à p* :

$$rg(X) = p.$$

Autrement dit, les colonnes de X sont linéairement indépendantes ; aucune colonne de X n'est combinaison linéaire des autres.

ii) *Les erreurs sont toutes de moyennes nulles* :

$$E(\mathcal{E}_i) = 0 \qquad i = 1, ..., n,$$

ce qui peut s'écrire

$$\mathcal{E}(\mathcal{E}) = \mathcal{E}\begin{pmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_n \end{pmatrix} = \begin{pmatrix} E(\mathcal{E}_1) \\ E(\mathcal{E}_2) \\ \vdots \\ E(\mathcal{E}_n) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = \mathcal{O}_{n\times 1}.$$

Ceci implique que $\mathcal{E}(Y) = \mathcal{E}(X\beta + \mathcal{E}) = X\beta + \mathcal{E}(\mathcal{E}) = X\beta$.

iii) *Les erreurs sont non corrélées et sont toutes de même variance* σ^2 :

$$cov(\mathcal{E}_i, \mathcal{E}_j) = 0 \qquad i, j = 1, ..., n; \quad i \neq j$$

et

$$var(\mathcal{E}_i) = \sigma^2 \qquad i = 1, ..., n,$$

ce qui peut s'écrire

$$\mathcal{E}(\mathcal{E}\mathcal{E}') = \sigma^2 I_n.$$

On définit la **matrice de dispersion** ou **matrice de variance-covariance** du vecteur Y de taille n par

$$D(Y) = Cov(Y, Y) = (cov(Y_i, Y_j))_{i,j=1,...,n}.$$

On note cette matrice Σ_Y.

L'hypothèse iii) implique que

$$\Sigma_Y = \sigma^2 I_n.$$

En effet,

$$\begin{aligned}
\Sigma_Y &= Cov(Y, Y) \\
&= \mathcal{E}[(Y - \mathcal{E}(Y))(Y - \mathcal{E}(Y))'] \\
&= \mathcal{E}[(X\beta + \mathcal{E} - X\beta)(X\beta + \mathcal{E} - X\beta)'] \\
&= \mathcal{E}[\mathcal{E}\mathcal{E}'] \\
&= \sigma^2 I_n.
\end{aligned}$$

◇

3.2 Estimateurs des coefficients de la régression

On cherche un estimateur $\hat{\beta}$ du vecteur de paramètres β afin de pouvoir estimer $\mathcal{E}(Y) = X\beta$ par $\hat{Y} = X\hat{\beta}$.

On souhaite minimiser les erreurs. On utilise le critère des moindres carrés, c'est-à-dire qu'on estime β par le vecteur $\hat{\beta}$ qui minimise

$$\begin{aligned} S(\beta) &= \|\mathcal{E}\|^2 \\ &= \|Y - X\beta\|^2 \\ &= (Y - X\beta)'(Y - X\beta) \\ &= Y'Y - 2Y'X\beta + \beta'X'X\beta. \end{aligned}$$

On obtient $\hat{\beta}$ en annulant la dérivée de $S(\beta)$ par rapport à β :

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'Y + 2X'X\beta = 0$$

$$\Leftrightarrow X'X\beta = X'Y.$$

Cet ensemble de p équations à p inconnues est appelé le **système d'équations normales.**

La matrice X'X est non singulière puisque, par hypothèse, X est de rang plein. Le système admet alors une solution unique :

$$\boxed{\hat{\beta} = (X'X)^{-1}X'Y.}$$

Propriétés de $\hat{\beta}$

1) $\hat{\beta}$ est un estimateur *linéaire* de β.

2) $\hat{\beta}$ est un estimateur *non biaisé* de β.

En effet,

$$\begin{aligned} \mathcal{E}(\hat{\beta}) &= \mathcal{E}((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'\mathcal{E}(Y) \\ &= (X'X)^{-1}X'X\beta \\ &= \beta. \end{aligned}$$

◇

3) $D(\hat{\beta}) = \sigma^2(X'X)^{-1}$.

En effet,

$$\begin{aligned} D(\hat{\beta}) &= D((X'X)^{-1}X'Y) \\ &= (X'X)^{-1}X'\Sigma_Y X(X'X)^{-1} \\ &= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}. \end{aligned}$$

◇

4) Théorème de Gauss-Markov

> Dans la classe des estimateurs linéaires sans biais de $l = a'\beta$ (où a est un vecteur de constantes de taille p), l'estimateur des moindres carrés $\hat{l} = a'\hat{\beta}$ est l'unique estimateur de variance minimum.
> On dira que $\hat{l} = a'\hat{\beta}$ est l'unique estimateur *BLUE* (Best Linear Unbiaised Estimator) de $l = a'\beta$.

Preuve

i) $\hat{l}$ est un estimateur linéaire sans biais de l :

$$E(\hat{l}) = E(a'\hat{\beta}) = a'\mathcal{E}(\hat{\beta}) = a'\beta = l$$

$$\hat{l} = a'\hat{\beta} = a'(X'X)^{-1}X'Y = c'Y$$

ii) $\hat{l}$ est l'unique estimateur BLUE de l :

Soit $\hat{t} = d'Y$ un autre estimateur linéaire non biaisé de l ($d' \neq c'$).

Montrons que $var(\hat{t}) > var(\hat{l}) = var(a'\hat{\beta}) = a'D(\hat{\beta})a = \sigma^2 a'(X'X)^{-1}a$.

$$E(\hat{t}) = E(d'Y) = d'\mathcal{E}(Y) = d'X\beta.$$

Comme $\hat{t}$ est non biaisé :

$$E(\hat{t}) = l = a'\beta.$$

Donc,

$$d'X = a'.$$

D'autre part,

$$\begin{aligned} cov(\hat{t},\hat{l}) &= cov(d'Y, c'Y) \\ &= E[(d'Y - d'X\beta)(c'Y - c'X\beta)] \\ &= d'\mathcal{E}[(Y - X\beta)(Y - X\beta)']c \\ &= d'\mathcal{E}(\mathcal{E}\mathcal{E}')c \\ &= \sigma^2 d'c \\ &= \sigma^2 d'X(X'X)^{-1}a \\ &= \sigma^2 a'(X'X)^{-1}a \\ &= var(\hat{l}). \end{aligned}$$

Donc,

$$\begin{aligned} var(\hat{t} - \hat{l}) &= var(\hat{t}) + var(\hat{l}) - 2cov(\hat{t},\hat{l}) \\ &= var(\hat{t}) - var(\hat{l}). \end{aligned}$$

Comme $\hat{t} \neq \hat{l}$:

$$var(\hat{t} - \hat{l}) > 0.$$

Finalement ,

$$var(\hat{t}) > var(\hat{l}).$$

◇

Conséquence

Si $\hat{l} = a'\hat{\beta}$ est l'unique estimateur BLUE de $l = a'\beta$,
alors $\hat{\beta}_j$ est l'unique estimateur BLUE de β_j $(j = 0, ..., p-1)$ et $\hat{\beta}$ est l'unique estimateur BLUE de β.

3.3 Estimateurs des coefficients de la régression dans le cas d'un seul régresseur

Supposons qu'on souhaite expliquer la variable aléatoire Y à l'aide d'un seul régresseur X par une relation linéaire

$$Y = \beta_0 + \beta_1 X + \mathcal{E}.$$

Si on mesure Y et X sur $E = \{1, ..., n\}$, on a n relations linéaires

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, ..., n$$

où $y_i = Y(i)$ et $x_i = X(i)$.

Cherchons les estimations des moindres carrés de β_0 et de β_1, c'est-à-dire les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ qui minimisent

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

On obtient $\hat{\beta}_0$ et $\hat{\beta}_1$ en annulant la dérivée de $S(\beta_0, \beta_1)$ respectivement par rapport à β_0 et à β_1 :

$$\begin{cases} \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases}$$

$$\Leftrightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i y_i = \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 \end{cases}$$

Ce système de 2 équations à 2 inconnues est appelé le **système d'équations normales**.

On obtient

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

et

$$\boxed{\hat{\beta}_1 = \frac{\frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x}\bar{y}}{\frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2} = \frac{s_{xy}}{s_x^2}.}$$

3.4 Estimateur de la variance des erreurs

On estime la variance des erreurs par

$$\boxed{\hat{\sigma}^2 = \frac{(Y-\hat{Y})'(Y-\hat{Y})}{n-p}.}$$

Propriétés de $\hat{\sigma}^2$

1) $\hat{\sigma}^2$ est un estimateur *non biaisé* de σ^2.

2) Sous l'hypothèse supplémentaire de normalité des erreurs, on peut montrer que dans la classe des estimateurs non biaisés de σ^2, $\hat{\sigma}^2$ est l'unique estimateur de variance minimum.
On dira que $\hat{\sigma}^2$ est l'unique estimateur *UMVU* (Uniformly Minimum Variance Unbiaised) ou encore estimateur optimal de σ^2.

Chapitre 4

Evaluation de la qualité de la régression et validation

Ce chapitre est également basé sur le cours de "Modèles Statistiques Linéaires" de A. Hardy [Hardy04].

4.1 Coefficient de détermination

4.1.1 Analyse de la variance

Définissons

- la somme des carrés des écarts totaux :

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

- la somme des carrés des écarts expliqués par la régression :

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2,$$

- la somme des carrés des écarts inexpliqués par la régression :

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

L'équation de l'analyse de la variance est alors

$$\boxed{SST = SSR + SSE.}$$

Remarque

L'estimateur de la variance des erreurs peut maintenant s'écrire

$$\hat{\sigma}^2 = \frac{(Y - \hat{Y})'(Y - \hat{Y})}{n - p} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n - p} = \frac{SSE}{n - p}.$$

4.1.2 Coefficient de détermination

Le coefficient de détermination est défini par

$$R = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\left(\sum_{i=1}^{n}(y_i - \bar{y})^2 \sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2\right)^{\frac{1}{2}}}.$$

On remarque que le coefficient de détermination est égal au coefficient de corrélation entre Y et $\hat{Y}$:

$$R = \frac{s_{y\hat{y}}}{s_y s_{\hat{y}}} = r_{y\hat{y}}.$$

On a donc

$$0 \leq R^2 \leq 1.$$

D'autre part, on peut montrer que

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST}.$$

Le coefficient de détermination constitue donc un indicateur de la qualité du modèle de régression ; il mesure la proportion de la variation totale de Y expliquée par la régression.

Plus la part de variation de Y expliquée par la régression est importante, plus R^2 s'approche de 1. (Si $R^2 = 1$, alors $y_i = \hat{y}_i \ \forall\, i$.)

4.2 Inférence sur les coefficients de la régression

4.2.1 Le modèle et ses hypothèses

Modèle :

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \mathcal{E}$$

Si on mesure ces p variables sur $E = \{1, ..., n\}$, on obtient n équations

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad i = 1, ..., n$$

qui peuvent s'écrire sous forme matricielle

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1,p-1} \\ 1 & x_{21} & \dots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{n,p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_n \end{pmatrix}$$

ou encore

$$Y = X\beta + \mathcal{E}.$$

Hypothèses :

On suppose à nouveau que $rg(X) = p$ et que les erreurs $\mathcal{E}_i$ sont non corrélées et sont toutes de moyenne nulle et de même variance σ^2.

On suppose de plus que les erreurs suivent une distribution normale :

$$\mathcal{E}_i \sim N(0, \sigma^2) \qquad i = 1, ..., n.$$

On peut donc écrire

$$\mathcal{E} \sim \mathcal{N}_n(\mathcal{O}_{n\times 1}, \sigma^2 I_n).$$

Ceci implique

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n).$$

Conséquences :

Sous l'hypothèse supplémentaire de normalité des erreurs, on peut montrer que

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X'X)^{-1});$$

$$\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p};$$

$$\hat{\beta} \amalg \hat{\sigma}^2.$$

4.2.2 Tests d'hypothèses

1. Test de Fisher-Snedecor

Introduction

Dans cette section, nous allons construire un test de l'hypothèse $H_0 : A\beta = C$ où

- A est une matrice connue de taille $(q \times p)$ et de rang q,
- C est un vecteur connu de dimension q.

Une telle hypothèse est appelée **hypothèse linéaire générale.**

Estimateurs des coefficients de la régression sous contraintes linéaires

Cherchons l'estimateur des moindres carrés de β sous la contrainte linéaire $A\beta = C$.

On cherche donc l'estimateur $\hat{\beta}_{H_0}$ de β qui minimise

$$\|\mathcal{E}\|^2 = \|Y - X\beta\|^2 = (Y - X\beta)'(Y - X\beta)$$

et qui vérifie $H_0 : A\beta = C$.

Pour trouver cet estimateur, on utilise la méthode des multiplicateurs de Lagrange.

La fonction à optimiser est $(Y - X\beta)'(Y - X\beta)$

Les contraintes linéaires sont

$$\begin{aligned} A\beta = C &\Leftrightarrow a_i'\beta = c_i \qquad && i = 1, ..., q \\ &\Leftrightarrow a_i'\beta - c_i = 0 && i = 1, ..., q \end{aligned}$$

où a_i' est la ligne i de la matrice A et c_i la composante i du vecteur C.

On considère donc l'expression

$$\sum_{i=1}^{q} \lambda_i (a_i'\beta - c_i) = \lambda'(A\beta - C) = (\beta' A' - C')\lambda.$$

On forme ensuite la fonction

$$\begin{aligned} F(\beta, \lambda) &= (Y - X\beta)'(Y - X\beta) + (\beta' A' - C')\lambda \\ &= Y'Y - 2\beta' X'Y + \beta' X'X\beta + \beta' A'\lambda - C'\lambda. \end{aligned}$$

Nous allons maintenant résoudre le système

$$\begin{cases} A\beta = C \\ \frac{\partial F(\beta,\lambda)}{\partial \beta} = 0 \end{cases}$$

Considérons tout d'abord la seconde équation :

$$\frac{\partial F(\beta, \lambda)}{\partial \beta} = 0 \Leftrightarrow -2X'Y + 2X'X\beta + A'\lambda = 0.$$

On obtient alors

$$\hat{\beta}_{H_0} = (X'X)^{-1}X'Y - \frac{1}{2}(X'X)^{-1}A'\hat{\lambda}_{H_0} = \hat{\beta} - \frac{1}{2}(X'X)^{-1}A'\hat{\lambda}_{H_0} \qquad (*).$$

En remplaçant β par $\hat{\beta}_{H_0}$ dans la première équation, on obtient

$$C = A\hat{\beta}_{H_0} = A\hat{\beta} - \frac{1}{2}A(X'X)^{-1}A'\hat{\lambda}_{H_0}$$

d'où on peut déduire

$$-\frac{1}{2}\hat{\lambda}_{H_0} = [A(X'X)^{-1}A']^{-1}(C - A\hat{\beta}).$$

En remplaçant $-\frac{1}{2}\hat{\lambda}_{H_0}$ dans $(*)$, on obtient finalement

$$\boxed{\hat{\beta}_{H_0} = \hat{\beta} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(C - A\hat{\beta}).}$$

Test de Fisher-Snedecor

On note

$$SSE = (Y - X\hat{\beta})'(Y - X\hat{\beta})$$

et

$$SSE_{H_0} = (Y - X\hat{\beta}_{H_0})'(Y - X\hat{\beta}_{H_0}).$$

Théorème

Sous $H_0 : A\beta = C$, on a $\dfrac{\frac{SSE_{H_0} - SSE}{q}}{\frac{SSE}{(n-p)}} \sim F_{q,n-p}$

Preuve

1) $\boxed{(SSE_{H_0} - SSE) \amalg SSE}$

On peut calculer

$$SSE_{H_0} - SSE = (A\hat{\beta} - C)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - C).$$

$(SSE_{H_0} - SSE)$ est donc une fonction de $\hat{\beta}$.

Comme $\hat{\beta} \amalg \hat{\sigma}^2 = \frac{SSE}{n-p}$,, on a

$$(SSE_{H_0} - SSE) \amalg SSE.$$

2) $\boxed{\text{Si } A\beta = C, \text{ on a } \quad \frac{SSE_{H_0} - SSE}{\sigma^2} \sim \chi^2_q}$

Comme

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2 (X'X)^{-1})$$

et A $(q \times p)$ est de rang q, on a

$$A\hat{\beta} \sim \mathcal{N}_q(A\beta, \sigma^2 A(X'X)^{-1}A').$$

Si $A\beta = C$, on a

$$A\hat{\beta} \sim \mathcal{N}_q(C, \sigma^2 A(X'X)^{-1}A')$$

et donc

$$(A\hat{\beta} - C)'(D(A\hat{\beta}))^{-1}(A\hat{\beta} - C) \sim \chi^2_q \qquad (*).$$

Remarquons que

$$D(A\hat{\beta}) = AD(\hat{\beta})A' = \sigma^2 A(X'X)^{-1}A'$$

et donc

$$(D(A\hat{\beta}))^{-1} = \frac{1}{\sigma^2}[A(X'X)^{-1}A']^{-1}.$$

En remplaçant $(D(A\hat{\beta}))^{-1}$ dans $(*)$: si $A\beta = C$, on a

$$\frac{(A\hat{\beta} - C)'[A(X'X)^{-1}A']^{-1}(A\hat{\beta} - C)}{\sigma^2} \sim \chi^2_q$$

c'est-à-dire

$$\frac{SSE_{H_0} - SSE}{\sigma^2} \sim \chi^2_q.$$

3) $\boxed{\frac{SSE}{\sigma^2} \sim \chi^2_{n-p}}$

En effet, on sait que $\frac{SSE}{\sigma^2} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

Conclusion

1), 2) et 3) impliquent que sous $H_0 : A\beta = C$, on a

$$\frac{\frac{SSE_{H_0} - SSE}{q}}{\frac{SSE}{(n-p)}} \sim F_{q,n-p}.$$

◇

Construction du test d'hypothèses de Fisher-Snedecor

On sait déjà que

$$E\left[\frac{SSE}{n-p}\right] = E\left[\hat{\sigma}^2\right] = \sigma^2.$$

On peut calculer

$$E\left[\frac{SSE_{H_0} - SSE}{q}\right] = \sigma^2 + \frac{(A\beta - C)'[A(X'X)^{-1}A']^{-1}(A\beta - C)}{q} = \sigma^2 + \delta.$$

$\delta \geq 0$ puisque

$$\delta = \frac{(A\beta - C)'[A(X'X)^{-1}A']^{-1}(A\beta - C)}{q}$$

et $A(X'X)^{-1}A'$ est définie positive.

Si $H_0 : A\beta = C$ est vraie,

alors

$$\delta = 0$$

et

$$E\left[\frac{SSE_{H_0} - SSE}{q}\right] = E\left[\frac{SSE}{n-p}\right] = \sigma^2.$$

On peut alors s'attendre à avoir

$$\boxed{\frac{\frac{SSE_{H_0} - SSE}{q}}{\frac{SSE}{(n-p)}} = 1.}$$

Si $H_0 : A\beta = C$ est fausse,

alors

$$\delta > 0$$

et

$$E\left[\frac{SSE_{H_0} - SSE}{q}\right] > E\left[\frac{SSE}{n-p}\right].$$

On peut dans ce cas s'attendre à ce que

$$\boxed{\frac{\frac{SSE_{H_0}-SSE}{q}}{\frac{SSE}{(n-p)}} > 1.}$$

Conclusion

On rejettera H_0 lorsque

$$\frac{\frac{SSE_{H_0}-SSE}{q}}{\frac{SSE}{(n-p)}}$$

est significativement grand.

La règle de décision du test de l'hypothèse nulle

$$H_0 : A\beta = C$$

contre l'hypothèse alternative

$$H_A : A\beta \neq C$$

sera la suivante :

On rejette H_0 au niveau de signification α si $F_{obs} > F_{1-\alpha,q,n-p}$.

Cas particulier

Supposons qu'on souhaite tester l'hypothèse nulle

$$H_0 : \beta_k = \beta_j = 0$$

contre l'hypothèse alternative

$$H_A : \beta_k \neq 0 \quad \text{ou} \quad \beta_j \neq 0.$$

L'hypothèse nulle peut encore s'écrire $H_0 : A\beta = \mathcal{O}_{q\times 1}$ où A est de rang $q = 2$.

Sous H_0, on a

$$\frac{\frac{SSE_{H_0} - SSE}{2}}{\frac{SSE}{n-p}} \sim F(2, n-p)$$

où on peut écrire

$$\begin{aligned} SSE_{H_0} - SSE &= SSE(x_1, ..., x_{k-1}, x_{k+1}, ..., x_{j-1}, x_{j+1}, ..., x_{p-1}) \\ &\quad - SSE(x_1, ..., x_{p-1}) \\ &= SSR(x_k, x_j | x_1, ..., x_{k-1}, x_{k+1}, ..., x_{j-1}, x_{j+1}, ..., x_{p-1}). \end{aligned}$$

Ce dernier terme peut encore s'écrire

$$SSR(x_k, x_j | x_1, ..., x_{k-1}, x_{k+1}, ..., x_{j-1}, x_{j+1}, ..., x_{p-1})$$

$$= SSR(x_k | x_1, ..., x_{k-1}, x_{k+1}, ..., x_{j-1}, x_{j+1}, ..., x_{p-1})$$

$$+ SSR(x_j | x_1, ..., x_{j-1}, x_{j+1}, ..., x_{p-1})$$

où

$$\begin{aligned} SSR(x_j | x_1, ..., x_{j-1}, x_{j+1}, ..., x_{p-1}) &= SSR(x_1, ..., x_{p-1}) \\ &\quad - SSR(x_1, ..., x_{j-1}, x_{j+1}, ..., x_{p-1}) \end{aligned}$$

est la contribution supplémentaire de x_j à SSR (somme des carrés des écarts expliqués par la régression) lorsque les autres régresseurs sont déjà dans le modèle.

2. Test de Student

Rejeter l'hypothèse nulle $H_0 : A\beta = C$ revient à affirmer que

$$\exists i \in \{1, .., q\} \quad \text{tel que} \quad a_i'\beta \neq c_i$$

où a_i' est la ligne i de la matrice A et c_i la composante i du vecteur C.

Si le test de Fisher-Snedecor conduit au rejet de $H_0 : A\beta = C$, on peut tester séparément chaque contrainte linéaire, c'est-à-dire chacune des q hypothèses nulles

$$H_{0i} : a_i'\beta = c_i \qquad i = 1, ..., q.$$

On cherche donc ici un test d'une hypothèse de la forme $H_0 : a'\beta = c$ où a est un vecteur de constantes de taille p et c est une constante.

On peut montrer que

Sous $H_0 : a'\beta = c$, on a $\dfrac{a'\hat{\beta} - c}{\left(a'(X'X)^{-1}a\frac{SSE}{n-p}\right)^{\frac{1}{2}}} \sim t_{n-p}$

Preuve

1) Si $a'\beta = c$, on a $U = \dfrac{a'\hat{\beta} - c}{\sigma(a'(X'X)^{-1}a)^{\frac{1}{2}}} \sim N(0,1)$

En effet, comme

$$\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X'X)^{-1}),$$

on a

$$a'\hat{\beta} \sim N(a'\beta, \sigma^2 a'(X'X)^{-1}a)$$

et donc, si $a'\beta = c$:

$$a'\hat{\beta} \sim N(c, \sigma^2 a'(X'X)^{-1}a).$$

On en déduit donc que, si $a'\beta = c$:

$$U = \frac{a'\hat{\beta} - c}{\sigma\,(a'(X'X)^{-1}a)^{\frac{1}{2}}} \sim N(0,1).$$

2) $\boxed{V = \frac{SSE}{\sigma^2} \sim \chi^2_{n-p}}$

3) $\boxed{U \amalg V}$

En effet, U est une fonction de $\hat{\beta}$, V est une fonction de $SSE = (n-p)\hat{\sigma}^2$, et $\hat{\beta} \amalg \hat{\sigma}^2$.

Conclusion

1), 2) et 3) impliquent que sous $H_0 : a'\beta = c$, on a

$$T = \frac{U}{\sqrt{\frac{V}{(n-p)}}} = \frac{a'\hat{\beta} - c}{\left(a'(X'X)^{-1}a\frac{SSE}{n-p}\right)^{\frac{1}{2}}} \sim t_{n-p}.$$

◇

Conclusion

La règle de décision du test de l'hypothèse nulle

$$H_0 : a'\beta = c$$

contre l'hypothèse alternative

$$H_A : a'\beta \neq c$$

sera la suivante :

$\boxed{\text{On rejette } H_0 \text{ au niveau de signification } \alpha \text{ si } |t_{obs}| > t_{1-\frac{\alpha}{2},n-p}.}$

Cas particulier

On se demande si la variable x_k a une influence sur la variable Y.

On souhaite donc tester l'hypothèse nulle

$$H_0 : \beta_k = 0$$

contre l'hypothèse alternative

$$H_A : \beta_k \neq 0.$$

L'hypothèse nulle peut encore s'écrire $H_0 : a'\beta = 0$ où $a' = (0, ..., 1, ..., 0)$.

Sous H_0, on a

$$T = \frac{\hat{\beta}_k}{\left((X'X)^{-1}_{kk}\frac{SSE}{n-p}\right)^{\frac{1}{2}}} = \frac{\hat{\beta}_k}{s_{\hat{\beta}_k}} \sim t_{n-p}$$

où $(X'X)^{-1}_{kk}\frac{SSE}{n-p} = \hat{\sigma}^2(X'X)^{-1}_{kk} = (\hat{D}(\hat{\beta}))_{kk} = \hat{var}(\hat{\beta}_k) = s^2_{\hat{\beta}_k}$.

On rejettera H_0 au niveau de signification α si $|t_{obs}| > t_{1-\frac{\alpha}{2},n-p}$.

3. Equivalence des tests de Fisher-Snedecor et de Student

Le test de Fisher-Snedecor et le test de Student sont équivalents :

sous $H_0 : a'\beta = c$, on a

$$\frac{\frac{SSE_{H_0}-SSE}{1}}{\frac{SSE}{n-p}} = \frac{(a'\hat{\beta}-c)^2}{a'(X'X)^{-1}a\frac{SSE}{n-p}} \sim F_{1,n-p}$$

$$\Leftrightarrow$$

$$\sqrt{\frac{\frac{SSE_{H_0}-SSE}{1}}{\frac{SSE}{n-p}}} = \frac{a'\hat{\beta}-c}{\left(a'(X'X)^{-1}a\frac{SSE}{n-p}\right)^{\frac{1}{2}}} \sim t_{n-p}.$$

4.2.3 Intervalle de confiance pour β_k

L'intervalle de confiance à $100(1-\alpha)\%$ du paramètre β_k $(k = 0, ..., p-1)$ est donné par

$$\boxed{I_\alpha(\beta_k) = \hat{\beta}_k \pm t_{1-\frac{\alpha}{2},n-p}s_{\hat{\beta}_k}.}$$

4.3 Intervalle de confiance pour $E(Y_0)$

Supposons que nous disposions du vecteur $x_0 = (1, x_{01}, ..., x_{0,p-1})'$, où les x_{0j} $(j = 1, ..., p-1)$ sont les valeurs des régresseurs pour un nouvel individu noté 0.

Nous cherchons un intervalle de confiance pour $E(Y_0) = x_0'\beta$.

On a

$$\boxed{I_\alpha(E(Y_0)) = \hat{y}_0 \pm t_{1-\frac{\alpha}{2},n-p} s_{\hat{y}_0},}$$

où $\hat{y}_0 = x_0'\hat{\beta}$ est l'estimation ponctuelle de $E(Y_0) = x_0'\beta$.

Troisième partie

La régression linéaire pour des données intervalles

Chapitre 5

Méthodologie

5.1 Le modèle

On souhaite expliquer une variable aléatoire Y *de type intervalle* à partir de $p-1$ régresseurs $X_1, ..., X_{p-1}$ *de type intervalle* par une relation linéaire

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \mathcal{E}$$

où $\mathcal{E}$ représente le terme d'erreur.

Soit $E = \{1, ..., n\}$ un ensemble de n objets sur lesquels on a mesuré ces p variables.

On a donc n relations linéaires :

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad i = 1, ..., n$$

où

$$y_i = Y(i) = [a_i, b_i] \ , \quad \varepsilon_i = [\varepsilon_{iL}, \varepsilon_{iU}] \qquad i = 1, ..., n$$

et

$$x_{ij} = X_j(i) = [c_{ij}, d_{ij}] \qquad i = 1, ..., n; j = 1, ..., p-1.$$

Ces n équations peuvent s'écrire sous forme matricielle :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & [c_{11}, d_{11}] & \dots & [c_{1,p-1}, d_{1,p-1}] \\ 1 & [c_{21}, d_{21}] & \dots & [c_{2,p-1}, d_{2,p-1}] \\ \vdots & \vdots & & \vdots \\ 1 & [c_{n1}, d_{n1}] & \dots & [c_{n,p-1}, d_{n,p-1}] \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_1 \\ \mathcal{E}_2 \\ \vdots \\ \mathcal{E}_n \end{pmatrix}$$

ou encore

$$Y = X\beta + \mathcal{E}$$

où

- Y est un vecteur aléatoire observable de taille n (dont les composantes sont des variables *intervalles*),
- X est la matrice de régression (connue) de taille $(n \times p)$,
- β est un vecteur de paramètres inconnus de taille p,
- $\mathcal{E}$ est un vecteur aléatoire inobservable de taille n (dont les composantes sont des variables *intervalles*).

5.2 La méthode du centre

La méthode du centre a été introduite par L. Billard et E. Diday dans [Billard00].

Cette méthode consiste à ajuster le modèle de régression linéaire classique sur le centre des intervalles.

On prédit alors les bornes inférieure et supérieure de la variable dépendante $y_k = [a_k, b_k]$ pour un objet k, respectivement par le minimum et le maximum des valeurs trouvées en appliquant ce modèle aux bornes inférieures et supérieures des variables indépendantes $x_{kj} = [c_{kj}, d_{kj}]$ $(j = 1, ..., p-1)$ mesurées sur cet objet.

Il s'agit de *la* méthode programmée dans le logiciel SODAS 2.

5.2.1 Le modèle

Hypothèse :

La méthode du centre suppose l'uniformité de la distribution à l'intérieur des intervalles.

Autrement dit, pour $i \in \{1, ..., n\}$, on suppose que les valeurs $x_{[Y]}, x_{[X_1]}, ..., x_{[X_{p-1}]}$ de tout vecteur de description individuel $x \in vir(d_i)$ sont uniformément distribuées dans les intervalles $[a_i, b_i]$, $[c_{i1}, d_{i1}]$,...,$[c_{i,p-1}, d_{i,p-1}]$ respectivement.

Modèle :

On considère les variables *quantitatives* Y_C et $X_{1C}, ..., X_{p-1,C}$ qui prennent comme valeur, respectivement, le centre des intervalles assumés par les variables Y et $X_1, ..., X_{p-1}$.

Le modèle correspondant est le suivant :

$$Y_C = \beta_{0C} + \beta_{1C} X_{1C} + ... + \beta_{p-1,C} X_{p-1,C} + \mathcal{E}_C.$$

Puisqu'on a mesuré Y et $X_1, ..., X_{p-1}$ sur $E = \{1, ..., n\}$, on obtient ici n relations linéaires

$$y_{iC} = \beta_{0C} + \beta_{1C} x_{i1C} + ... + \beta_{p-1,C} x_{i,p-1,C} + \varepsilon_{iC} \qquad i = 1, ..., n$$

où $y_{iC} = Y_C(i) = \frac{a_i + b_i}{2}$, $x_{ijC} = x_{jC}(i) = \frac{c_{ij} + d_{ij}}{2}$ et $\varepsilon_{iC} = \frac{\varepsilon_{iL} + \varepsilon_{iU}}{2}$.

Ces n équations peuvent s'écrire sous forme matricielle

$$\begin{pmatrix} Y_{1C} \\ Y_{2C} \\ \vdots \\ Y_{nC} \end{pmatrix} = \begin{pmatrix} 1 & x_{11C} & \dots & x_{1,p-1,C} \\ 1 & x_{21C} & \dots & x_{2,p-1,C} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1C} & \dots & x_{n,p-1,C} \end{pmatrix} \begin{pmatrix} \beta_{0C} \\ \beta_{1C} \\ \vdots \\ \beta_{p-1,C} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_{1C} \\ \mathcal{E}_{2C} \\ \vdots \\ \mathcal{E}_{nC} \end{pmatrix}$$

ou encore

$$Y_C = X_C \beta_C + \mathcal{E}_C.$$

5.2.2 Estimateurs des coefficients de la régression

L'estimateur des moindres carrés de β_C est donné par

$$\boxed{\hat{\beta}_C = (X_C' X_C)^{-1} X_C' Y_C.}$$

5.2.3 Estimateurs des coefficients de la régression dans le cas d'un seul régresseur

Supposons qu'on souhaite expliquer la variable aléatoire Y *de type intervalle* à l'aide d'un seul régresseur X *de type intervalle* par une relation linéaire

$$Y = \beta_0 + \beta_1 X + \mathcal{E}\,.$$

Si on mesure Y et X sur $E = \{1, ..., n\}$, on a n relations linéaires

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, ..., n$$

où $y_i = Y(i) = [a_i, b_i]$, $x_i = X(i) = [c_i, d_i]$ et $\varepsilon_i = [\varepsilon_{iL}, \varepsilon_{iU}]$.

Rappelons que dans le cas de variables classiques Y et X, les estimations des moindres carrés de β_0 et de β_1 sont données par

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

et

$$\boxed{\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}.}$$

Si Y et X sont des variables de type intervalle, on utilise les mêmes estimations en remplaçant les moyennes $\bar{y}$ et $\bar{x}$, et la covariance s_{xy}, par leurs équivalents symboliques définis dans le chapitre 2 :

$$\bar{y} = \frac{1}{2n}\sum_{i=1}^{n}(a_i + b_i)\,, \quad \bar{x} = \frac{1}{2n}\sum_{i=1}^{n}(c_i + d_i)\,,$$

$$s_{xy} = \frac{1}{4n}\sum_{i=1}^{n}(c_i + d_i)(a_i + b_i) - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(c_i + d_i)\right]\left[\sum_{i=1}^{n}(a_i + b_i)\right]\,,$$

et la variance s_x^2 par son équivalent symbolique :

$$s_x^2 = \frac{1}{4n}\sum_{i=1}^{n}(c_i + d_i)^2 - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(c_i + d_i)\right]^2 .$$

NB : Nous faisons donc une régression linéaire simple classique avec les centres des intervalles.

5.2.4 Prédiction de Y pour un nouvel objet

Supposons que nous disposions des valeurs (intervalles) $x_{k1}, ..., x_{k,p-1}$ des régresseurs $X_1, ..., X_{p-1}$ pour un nouvel objet $k \notin E$, où $x_{kj} = [c_{kj}, d_{kj}]$, $j = 1, ..., p-1$.

Nous prédisons alors les bornes inférieure et supérieure de $y_k = [a_k, b_k]$ respectivement par

$$\hat{a}_k = \min\{c_k'\hat{\beta}_C,\ d_k'\hat{\beta}_C\}$$

et

$$\hat{b}_k = \max\{c_k'\hat{\beta}_C,\ d_k'\hat{\beta}_C\}$$

où $c_k' = (1, c_{k1}, ..., c_{k,p-1})$ et $d_k' = (1, d_{k1}, ..., d_{k,p-1})$.

5.2.5 Exemples

Exemple 5.2.1

Nous avons imaginé le tableau de données suivant afin d'illustrer la méthode du centre :

k	Y : poids	X_1 : âge	X_2 : taille
1	[58, 66]	[18, 22]	[160, 168]
2	[60, 70]	[25, 35]	[158, 174]
3	[62, 72]	[36, 44]	[163, 173]
4	[65, 77]	[42, 58]	[156, 168]
5	[68, 80]	[54, 66]	[169, 193]

Régression linéaire simple

Nous nous intéressons tout d'abord au modèle

$$Y = \beta_0 + \beta_1 X_1 + \mathcal{E}_1.$$

On calcule

$$\bar{y} = 67.8\,; \qquad \bar{x}_1 = 40\,;$$

$$s^2_{x_1} = 250\,; \qquad s_{x_1 y} = 75\,.$$

On obtient alors

$$\hat{\beta}_{1C} = \frac{s_{x_1 y}}{s^2_{x_1}} = \frac{75}{250} = 0.3$$

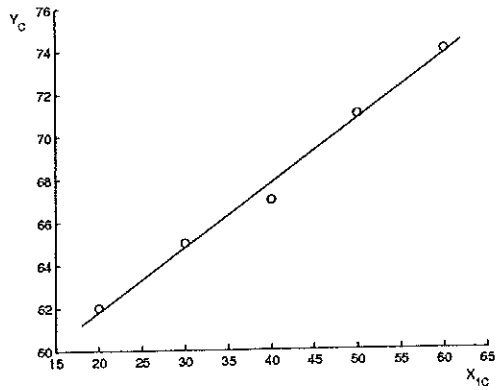
et

$$\hat{\beta}_{0C} = \bar{y} - \hat{\beta}_1 \bar{x}_1 = 67.8 - 0.3 \cdot 40 = 55.8\,.$$

On a donc

$$\boxed{\hat{Y}_C = 55.8 + 0.3\, X_{1C}\ .}$$

Nous avons représenté cette droite de régression estimée à partir des centres des intervalles de Y et de X_1 :



Si l'âge d'un objet k est donné par l'intervalle $x_{k1} = [20, 25]$, on prédit son poids par

$$\hat{y}_k = [61.8, 63.3]$$

en calculant

$$55.8 + 0.3 \cdot 20 = 61.8;$$

$$55.8 + 0.3 \cdot 25 = 63.3.$$

Régression linéaire multiple

Considérons maintenant le modèle contenant les deux régresseurs :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}.$$

On utilise les centres des intervalles, donc la matrice de régression

$$X_C = \begin{pmatrix} 1 & 20 & 164 \\ 1 & 30 & 166 \\ 1 & 40 & 168 \\ 1 & 50 & 172 \\ 1 & 60 & 176 \end{pmatrix}$$

et le vecteur $y_C = (62, 65, 67, 71, 74)'$ classiques.

En calculant $\hat{\beta}_C = (X_C' X_C)^{-1} X_C' y_C$, on obtient

$$\boxed{\hat{Y}_C = -11.571 + 0.171\, X_{1C} + 0.429\, X_{2C}.}$$

Si l'âge d'un objet k est donné par $x_{k1} = [20, 25]$ et sa taille par $x_{k2} = [165, 175]$, on prédit son poids par

$$\hat{y}_k = [62.63, 67.78]$$

en calculant

$$-11.571 + 0.171 \cdot 20 + 0.429 \cdot 165 = 62.63;$$

$$-11.571 + 0.171 \cdot 25 + 0.429 \cdot 175 = 67.78.$$

□

Exemple 5.2.2

Nous étudions les données abordées par L. Billard et E. Diday dans [Billard00] et issues de [Raju97].

On considère les variables suivantes :

Y : fréquence cardiaque (nombre de battements cardiaques par minute),
X_1 : pression artérielle systolique (en mm Hg),
X_2 : pression artérielle diastolique (en mm Hg).

Le tableau suivant contient les valeurs (intervalles) de la fréquence cardiaque Y et des pressions artérielles systolique X_1 et diastolique X_2 de 10 patients :

k	Y	X_1	X_2
1	[44, 68]	[90, 110]	[50, 70]
2	[60, 72]	[90, 130]	[70, 90]
3	[56, 90]	[140, 180]	[90, 100]
4	[70, 112]	[110, 142]	[80, 108]
5	[54, 72]	[90, 100]	[50, 70]
6	[70, 100]	[134, 142]	[80, 110]
7	[72, 100]	[130, 160]	[76, 90]
8	[76, 98]	[110, 190]	[70, 110]
9	[86, 96]	[138, 180]	[90, 110]
10	[86, 100]	[110, 150]	[78, 100]

Premier modèle de régression linéaire simple

Nous nous intéressons tout d'abord au modèle

$$Y = \beta_0 + \beta_1 X_1 + \mathcal{E}_1.$$

Dans les exemples 2.1.4 et 2.2.2 du chapitre 2, nous avions calculé

$$\bar{y} = 79.1\,; \qquad \bar{x}_1 = 131.3\,.$$

On calcule aussi

$$s^2_{x_1} = 495.41\,; \qquad s_{x_1 y} = 194.17\,.$$

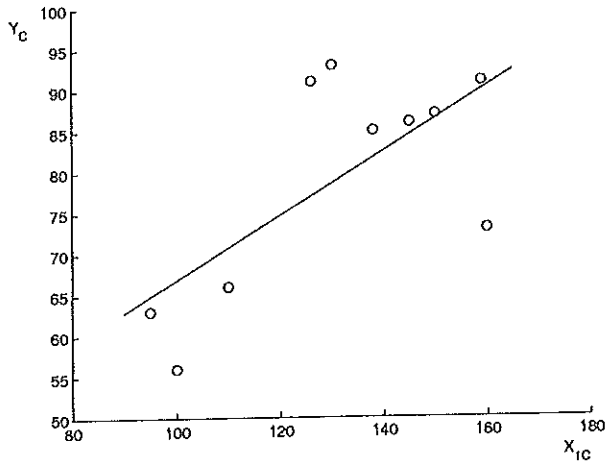
On estime alors les paramètres par

$$\hat{\beta}_{1C} = 0.392 \quad \text{et} \quad \hat{\beta}_{0C} = 27.630.$$

On a donc

$$\boxed{\hat{Y}_C = 27.630 + 0.392\, X_{1C}.}$$

Nous avons représenté cette droite de régression estimée à partir des centres des intervalles de Y et de X_1 :



Supposons maintenant que nous ayons mesuré la pression artérielle systolique d'un nouvel individu $k \notin E = \{1, ..., 10\}$: $x_{k1} = [118, 126]$. On prédit alors les bornes inférieure et supérieure de $y_k = [a_k, b_k]$ par

$$\hat{a}_k = 27.630 + 0.392 \cdot 118 = 73.89 \quad \text{et} \quad \hat{b}_k = 27.630 + 0.392 \cdot 126 = 77.02.$$

On estime donc la fréquence cardiaque de l'individu k par

$$\hat{y}_k = [73.89, 77.02].$$

Second modèle de régression linéaire simple

Nous nous intéressons ici au modèle

$$Y = \beta_0 + \beta_2 X_2 + \mathcal{E}_2.$$

Dans les exemples 2.1.4 et 2.2.2 du chapitre 2, nous avions calculé

$$\bar{y} = 79.1\,; \qquad \bar{x}_2 = 84.6.$$

Calculons aussi

$$s_{x_2}^2 = 182.44\,; \qquad s_{x_2 y} = 141.04.$$

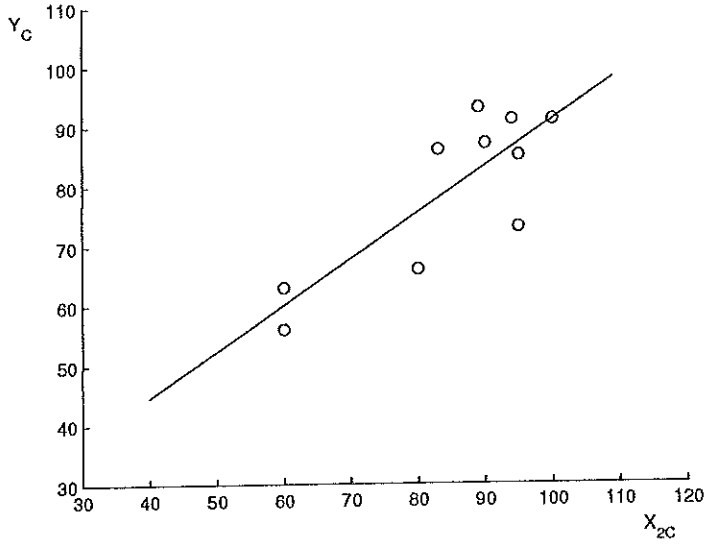
On obtient alors

$$\hat{\beta}_{2C} = 0.773 \quad \text{et} \quad \hat{\beta}_{0C} = 13.704.$$

On a donc

$$\boxed{\hat{Y}_C = 13.704 + 0.773\, X_{2C}.}$$

Nous avons représenté cette droite de régression estimée à partir des centres des intervalles de Y et de X_2 :



Supposons maintenant que nous ayons mesuré la pression artérielle diastolique de l'individu k : $x_{k2} = [85, 110]$. On prédit alors les bornes inférieure et supérieure de $y_k = [a_k, b_k]$ par

$$\hat{a}_k = 13.704 + 0.773 \cdot 85 = 79.41 \quad \text{et} \quad \hat{b}_k = 13.704 + 0.773 \cdot 110 = 98.73.$$

On estime donc la fréquence cardiaque de l'individu k par

$$\hat{y}_k = [79.41, 98.73].$$

Modèle de régression linéaire multiple

Nous considérons maintenant le modèle incluant les 2 régresseurs X_1 et X_2 :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}.$$

On calcule

$$\hat{\beta}_C = \left(X_C' X_C\right)^{-1} X_C' y_C$$

où

$$X_C = \begin{pmatrix} 1 & 100 & 60 \\ 1 & 110 & 80 \\ 1 & 160 & 95 \\ 1 & 126 & 94 \\ 1 & 95 & 60 \\ 1 & 138 & 95 \\ 1 & 145 & 83 \\ 1 & 150 & 90 \\ 1 & 159 & 100 \\ 1 & 130 & 89 \end{pmatrix} \quad \text{et} \quad y_C = \begin{pmatrix} 56 \\ 66 \\ 73 \\ 91 \\ 63 \\ 85 \\ 86 \\ 87 \\ 91 \\ 93 \end{pmatrix}.$$

On obtient

$$\hat{\beta}_C = \begin{pmatrix} \hat{\beta}_{0C} \\ \hat{\beta}_{1C} \\ \hat{\beta}_{2C} \end{pmatrix} = \begin{pmatrix} 14.165 \\ -0.040 \\ 0.830 \end{pmatrix}.$$

On a donc

$$\boxed{\hat{Y}_C = 14.165 - 0.040\, X_{1C} + 0.830\, X_{2C}.}$$

Les pressions artérielles systolique et diastolique de l'individu k étant respectivement $x_{k1} = [118, 126]$ et $x_{k2} = [85, 110]$, on prédit les bornes inférieure et supérieure de $y_k = [a_k, b_k]$ par

$$\hat{a}_k = 14.165 - 0.040 \cdot 118 + 0.830 \cdot 85 = 80.00$$

et

$$\hat{b}_k = 14.165 - 0.040 \cdot 126 + 0.830 \cdot 110 = 100.43.$$

On estime donc la fréquence cardiaque de l'individu k par

$$\hat{y}_k = [80.00, 100.43].$$

□

5.2.6 Remarque : les méthodes de la borne inférieure et de la borne supérieure

Les méthodes de la borne inférieure et de la borne supérieure sont illustrées par L. Billard et E. Diday dans [Billard00].

Elles sont analogues à la méthode du centre.

La méthode de la borne inférieure (respectivement supérieure) consiste à ajuster le modèle de régression linéaire classique sur les bornes inférieures (respectivement supérieures) des intervalles.

On prédit alors les bornes inférieure et supérieure de la variable dépendante y_k pour un objet k, respectivement par le minimum et le maximum des valeurs trouvées en appliquant ce modèle aux bornes inférieures et supérieures des variables explicatives x_{kj} $(j = 1, ..., p-1)$ mesurées sur cet objet.

On peut également appliquer simultanément ces deux méthodes et estimer ensuite la borne inférieure (respectivement supérieure) de y_k par celle estimée dans la méthode de la borne inférieure (respectivement supérieure).

NB : Dans ce cas, il n'est pas forcément assuré que la borne inférieure de l'intervalle prédit est inférieure à sa borne supérieure.

Exemple

Appliquons les méthodes de la borne inférieure et de la borne supérieure aux données artificielles de l'exemple 5.2.1.

Nous considérons le modèle de régression linéaire multiple :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}.$$

Méthode de la borne inférieure

Nous utilisons les bornes inférieures des intervalles, donc la matrice de régression

$$X_L = \begin{pmatrix} 1 & 18 & 160 \\ 1 & 25 & 158 \\ 1 & 36 & 163 \\ 1 & 42 & 156 \\ 1 & 54 & 169 \end{pmatrix}$$

et le vecteur $y_L = (58, 60, 62, 65, 68)'$ classiques.

En calculant $\hat{\beta}_L = (X_L' X_L)^{-1} X_L' y_L$, on obtient

$$\boxed{\hat{Y}_L = 61.565 + 0.291\, X_{1L} - 0.057\, X_{2L}.}$$

Si l'âge et la taille d'un objet k sont respectivement donnés par $x_{k1} = [20, 25]$ et $x_{k2} = [165, 175]$, on prédit son poids y_k par

$$\hat{y}_k = [57.98, 58.87]$$

en calculant

$$61.565 + 0.291 \cdot 20 - 0.057 \cdot 165 = 57.98\,;$$
$$61.565 + 0.291 \cdot 25 - 0.057 \cdot 175 = 58.87\,.$$

Méthode de la borne supérieure

On utilise les bornes supérieures des intervalles, donc la matrice de régression

$$X_U = \begin{pmatrix} 1 & 22 & 168 \\ 1 & 35 & 174 \\ 1 & 44 & 173 \\ 1 & 58 & 168 \\ 1 & 66 & 193 \end{pmatrix}$$

et le vecteur $y_U = (66, 70, 72, 77, 80)'$ classiques.

En calculant $\hat{\beta}_U = (X_U' X_U)^{-1} X_U' y_U$, on obtient

$$\boxed{\hat{Y}_U = 54.509 + 0.305\, X_{1U} + 0.027\, X_{2U}.}$$

L'âge et la taille de l'objet k étant respectivement donnés par $x_{k1} = [20, 25]$ et $x_{k2} = [165, 175]$, on prédit son poids y_k par

$$\hat{y}_k = [65.06, 66.86]$$

en calculant

$$54.509 + 0.305 \cdot 20 + 0.027 \cdot 165 = 65.06\,;$$
$$54.509 + 0.305 \cdot 25 + 0.027 \cdot 175 = 66.86\,.$$

Remarque

Si on a appliqué ces deux méthodes, on peut ensuite prédire la borne inférieure (respectivement supérieure) de y_k par celle estimée dans la méthode de la borne inférieure (respectivement supérieure) :

$$\hat{y}_k = [57.98, 66.86].$$

□

Comparaison avec la méthode du centre

Nous allons appliquer les méthodes de la borne inférieure et de la borne supérieure aux données réelles de l'exemple 5.2.2.

Nous comparerons ensuite les prédictions obtenues avec celles obtenues en utilisant la méthode du centre.

Cette comparaison a été faite par L. Billard et E. Diday dans [Billard00] uniquement en ce qui concerne le premier modèle de régression linéaire simple.

Méthode de la borne inférieure

Premier modèle de régression linéaire simple : $Y = \beta_0 + \beta_1 X_1 + \mathcal{E}_1$

On calcule

$$\bar{y}_L = 67.4, \qquad \bar{x}_{1L} = 114.2,$$

$$s^2_{x_{1L}} = 409.289, \qquad s_{x_{1L}y_L} = 135.244.$$

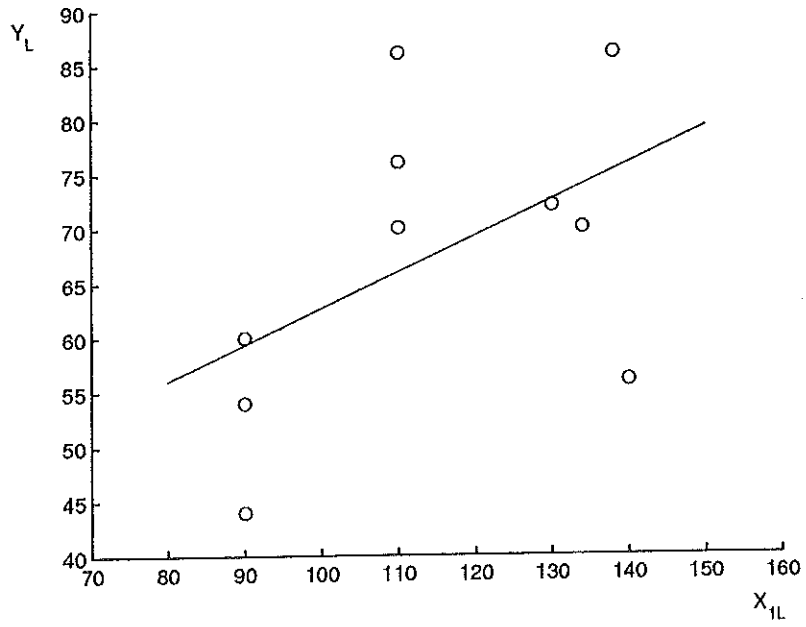
On obtient alors

$$\hat{\beta}_{1L} = 0.330 \quad \text{et} \quad \hat{\beta}_{0L} = 29.664.$$

Le modèle est donc

$$\boxed{\hat{Y}_L = 29.664 + 0.330\, X_{1L}.}$$

Nous avons représenté cette droite de régression estimée à partir des bornes inférieures des intervalles de Y et de X_1 :



Si la pression artérielle systolique d'un individu k est $x_{k1} = [118, 126]$, on prédit sa fréquence cardiaque par

$$\hat{y}_k = [68.60, 71.24].$$

Second modèle de régression linéaire simple : $Y = \beta_0 + \beta_2 X_2 + \mathcal{E}_2$

On a $\bar{y}_L = 67.4$. On calcule aussi

$$\bar{x}_{2L} = 73.4, \quad s^2_{x_{2L}} = 198.267 \quad \text{et} \quad s_{x_{2L}y_L} = 123.156.$$

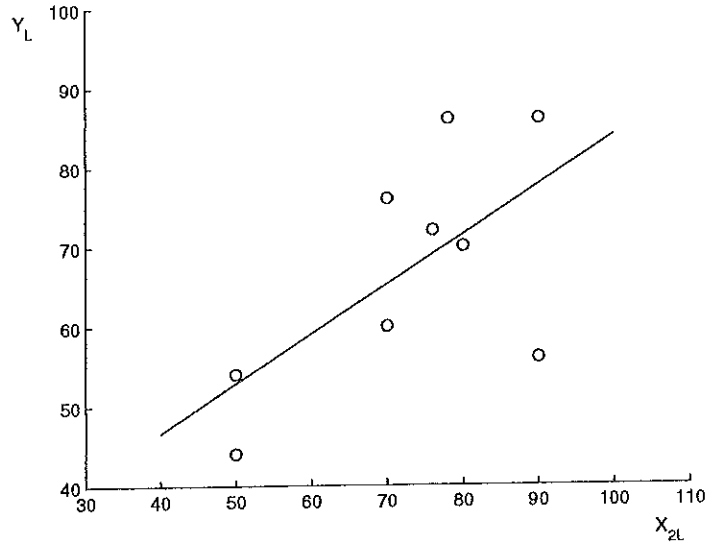
On obtient alors

$$\hat{\beta}_{2L} = 0.621 \quad \text{et} \quad \hat{\beta}_{0L} = 21.807.$$

Le modèle est donc

$$\boxed{\hat{Y}_L = 21.807 + 0.621\, X_{2L}.}$$

Nous avons représenté cette droite de régression estimée à partir des bornes inférieures des intervalles de Y et de X_2 :



La prédiction de la fréquence cardiaque de l'individu k, si sa pression artérielle diastolique est donnée par $x_{k2} = [85, 110]$, est alors

$$\hat{y}_k = [74.59, 90.12].$$

Modèle de régression linéaire multiple : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}$

On calcule

$$\hat{\beta}_L = \left(X_L' X_L\right)^{-1} X_L' y_L$$

où

$$X_L = \begin{pmatrix} 1 & 90 & 50 \\ 1 & 90 & 70 \\ 1 & 140 & 90 \\ 1 & 110 & 80 \\ 1 & 90 & 50 \\ 1 & 134 & 80 \\ 1 & 130 & 76 \\ 1 & 110 & 70 \\ 1 & 138 & 90 \\ 1 & 110 & 78 \end{pmatrix} \quad \text{et} \quad y_L = \begin{pmatrix} 44 \\ 60 \\ 56 \\ 70 \\ 54 \\ 70 \\ 72 \\ 76 \\ 86 \\ 86 \end{pmatrix}.$$

On obtient

$$\hat{\beta}_L = \begin{pmatrix} \hat{\beta}_{0L} \\ \hat{\beta}_{1L} \\ \hat{\beta}_{2L} \end{pmatrix} = \begin{pmatrix} 25.050 \\ -0.132 \\ 0.782 \end{pmatrix}.$$

On a donc

$$\boxed{\hat{Y}_L = 25.050 - 0.132\, X_{1L} + 0.782\, X_{2L}.}$$

La prédiction de la fréquence cardiaque de l'individu k, sachant que sa pression artérielle systolique est donnée par $x_{k1} = [118, 126]$ et diastolique par $x_{k2} = [85, 110]$, est alors

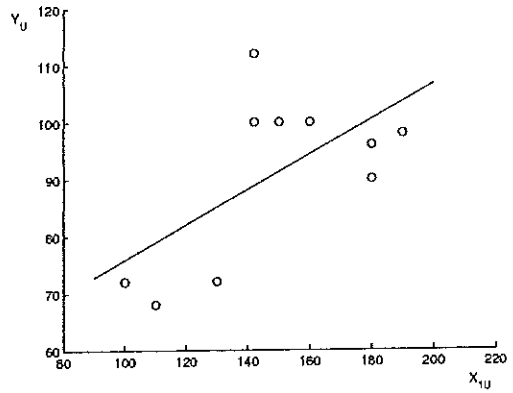
$$\hat{y}_k = [75.94, 94.44].$$

Méthode de la borne supérieure

En utilisant la méthode de la borne supérieure, on obtient les résultats suivants :

Premier modèle de régression linéaire simple :

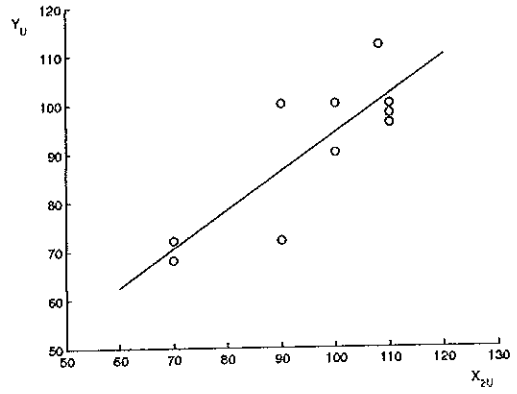
$$\boxed{\hat{Y}_U = 45.070 + 0.308\, X_{1U}.}$$



Si $x_{k1} = [118, 126]$, on estime $\hat{y}_k = [81.41, 83.88]$.

Second modèle de régression linéaire simple :

$$\boxed{\hat{Y}_U = 15.057 + 0.791\, X_{2U}.}$$



Si $x_{k2} = [85, 110]$, on estime $\hat{y}_k = [82.29, 102.07]$.

Modèle de régression linéaire multiple :

$$\boxed{\hat{Y}_U = 15.200 - 0.033\, X_{1U} + 0.840\, X_{2U}.}$$

Si $x_{k1} = [118, 126]$ et $x_{k2} = [85, 110]$, on estime $\hat{y}_k = [82.71, 103.44]$.

Remarque

Si on applique simultanément ces deux méthodes, on peut ensuite estimer la borne inférieure (respectivement supérieure) de y_k par celle estimée dans la méthode de la borne inférieure (respectivement supérieure).

Comparaison avec la méthode du centre

Nous avons repris dans le tableau suivant les prédictions de la fréquence cardiaque $y_k = [a_k, b_k]$ obtenues en utilisant la méthode du centre et les méthodes de la borne inférieure et de la borne supérieure, lorsque les pressions artérielles systolique et diastolique sont respectivement $x_{k1} = [118, 126]$ et $x_{k2} = [85, 110]$.

régresseurs	centre	borne inf.	borne sup.
X_1	[73.89, 77.02]	[68.60, 71.24]	[81.41, 83.88]
X_2	[79.41, 98.73]	[74.59, 90.12]	[82.29, 102.07]
X_1, X_2	[80.00, 100.43]	[75.94, 94.44]	[82.71, 103.44]

Considérons le premier modèle de régression $Y = \beta_0 + \beta_1 X_1 + \mathcal{E}_1$.

La borne inférieure de l'intervalle prédit par la méhode du centre est encadrée par les bornes inférieures des intervalles prédits par les deux autres méthodes :

$$68.60 < 73.89 < 81.41.$$

De même, la borne supérieure de l'intervalle prédit par la méthode du centre est encadrée par les bornes supérieures des intervalles prédits par les deux autres méthodes :

$$71.24 < 77.02 < 83.88.$$

Lorsqu'on applique simultanément les méthodes de la borne inférieure et de la borne supérieure, on peut estimer $\hat{y}_k = [68.60, 83.88]$.

Cet intervalle est alors beaucoup plus large que celui prédit par la méthode du centre, [73.89, 77.02].

Nous pouvons faire les mêmes comparaisons en ce qui concerne les deux autres modèles de régression.

La méthode du centre semble donc préférable.

□

5.3 La méthode du centre et de l'étendue

Cette méthode a été présentée par F. A. T. De Carvalho, E. A. Lima Neto et C. P. Tenorio dans [DeCarvalho04a].

Nous considérons deux approches.

La première consiste à ajuster deux modèles de régression linéaire classique. Dans le premier modèle, l'estimation du centre d'un intervalle assumé par la variable dépendante Y est basée sur les centres des intervalles assumés par les régresseurs X_j $(j = 1, ..., p-1)$.
Dans le second modèle, l'estimation de l'étendue d'un intervalle assumé par la variable dépendante est basée sur les étendues des intervalles assumés par les régresseurs.

Dans la seconde approche, on ajuste à nouveau deux modèles de régression linéaire classique, le premier nous permettant d'estimer le centre d'un intervalle assumé par la variable dépendante Y, et le second son étendue. A la différence de la première approche, chacune de ces estimations est maintenant basée à la fois sur les centres et les étendues des intervalles assumés par les régresseurs X_j $(j = 1, ..., p-1)$.

Dans chacune de ces deux approches, on estime les bornes inférieure et supérieure de la variable dépendante y_k pour un objet k à partir du centre et de l'étendue de celle-ci, estimés respectivement en appliquant le premier modèle de régression aux centres des variables indépendantes x_{kj} $(j = 1, ..., p-1)$ mesurées sur cet objet, et en appliquant le second modèle de régression aux étendues des x_{kj}.

5.3.1 Première méthode du centre et de l'étendue

Premier modèle

Il s'agit du modèle établi dans la méthode du centre :

$$Y_C = X_C \beta_C + \mathcal{E}_C.$$

Nous avions pour l'estimateur des moindres carrés de β_C

$$\boxed{\hat{\beta}_C = (X_C' X_C)^{-1} X_C' Y_C.}$$

On estime le centre de $y_k = [a_k, b_k]$ par

$$\hat{y}_{kC} = x'_{kC}\hat{\beta}_C$$

où $x'_{kC} = (1, \frac{c_{k,1}+d_{k,1}}{2}, ..., \frac{c_{k,p-1}+d_{k,p-1}}{2})$.

Second modèle

On considère les variables *quantitatives* Y_R et $X_{1R}, ..., X_{p-1,R}$ qui prennent comme valeur, respectivement, l'étendue des intervalles assumés par les variables Y et $X_1, ..., X_{p-1}$.

Le modèle correspondant est le suivant :

$$Y_R = \beta_{0R} + \beta_{1R}X_{1R} + ... + \beta_{p-1,R}X_{p-1,R} + \mathcal{E}_R.$$

Puisqu'on a mesuré Y et $X_1, ..., X_{p-1}$ sur $E = \{1, ..., n\}$, on obtient ici n relations linéaires

$$y_{iR} = \beta_{0R} + \beta_{1R}x_{i1R} + ... + \beta_{p-1,R}x_{i,p-1,R} + \varepsilon_{iR} \qquad i = 1, ..., n$$

où $y_{iR} = Y_R(i) = b_i - a_i$, $x_{ijR} = x_{jR}(i) = d_{ij} - c_{ij}$ et $\varepsilon_{iR} = \varepsilon_{iU} - \varepsilon_{iL}$.

Ces n équations peuvent s'écrire sous forme matricielle

$$\begin{pmatrix} Y_{1R} \\ Y_{2R} \\ \vdots \\ Y_{nR} \end{pmatrix} = \begin{pmatrix} 1 & x_{11R} & \cdots & x_{1,p-1,R} \\ 1 & x_{21R} & \cdots & x_{2,p-1,R} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1R} & \cdots & x_{n,p-1,R} \end{pmatrix} \begin{pmatrix} \beta_{0R} \\ \beta_{1R} \\ \vdots \\ \beta_{p-1,R} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_{1R} \\ \mathcal{E}_{2R} \\ \vdots \\ \mathcal{E}_{nR} \end{pmatrix}$$

ou encore

$$Y_R = X_R\beta_R + \mathcal{E}_R.$$

L'estimateur des moindres carrés de β_R est donné par

$$\boxed{\hat{\beta}_R = (X'_R X_R)^{-1} X'_R Y_R.}$$

On estime l'étendue de $y_k = [a_k, b_k]$ par

$$\hat{y}_{kR} = x'_{kR}\hat{\beta}_R$$

où $x'_{kR} = (1, d_{k1} - c_{k1}, ..., d_{k,p-1} - c_{k,p-1})$.

5.3.2 Seconde méthode du centre et de l'étendue

On considère les variables *quantitatives* $Y_C, X_{1C}, ..., X_{p-1,C}$ et $Y_R, X_{1R}, ..., X_{p-1,R}$ qui prennent comme valeur, respectivement, les centres et les étendues des intervalles assumés par $Y, X_1, ..., X_{p-1}$.

Premier modèle

Le modèle est le suivant :

$$\begin{aligned} Y_C = {} & \beta_{0C} + \beta_{1C}X_{1C} + ... + \beta_{p-1,C}X_{p-1,C} \\ & + \beta_{p,C}X_{1R} + ... + \beta_{2p-2,C}X_{p-1,R} + \mathcal{E}_C. \end{aligned}$$

Puisqu'on a mesuré Y et $X_1, ..., X_{p-1}$ sur $E = \{1, ..., n\}$, on obtient ici n relations linéaires

$$\begin{aligned} y_{iC} = {} & \beta_{0C} + \beta_{1C}x_{i1C} + ... + \beta_{p-1,C}x_{i,p-1,C} \\ & + \beta_{p,C}x_{i1R} + ... + \beta_{2p-2,C}x_{i,p-1,R} + \varepsilon_{iC} \qquad i = 1, ..., n \end{aligned}$$

où

$$y_{iC} = Y_C(i) = \frac{a_i + b_i}{2}, \qquad \varepsilon_{iC} = \frac{\varepsilon_{iL} + \varepsilon_{iU}}{2},$$

$$x_{ijC} = x_{jC}(i) = \frac{c_{ij} + d_{ij}}{2}, \qquad x_{ijR} = x_{jR}(i) = d_{ij} - c_{ij}.$$

Ces n équations peuvent s'écrire sous forme matricielle

$$\begin{pmatrix} Y_{1C} \\ Y_{2C} \\ \vdots \\ Y_{nC} \end{pmatrix} = \begin{pmatrix} 1 & x_{11C} & \dots & x_{1,p-1,C} & x_{11R} & \dots & x_{1,p-1,R} \\ 1 & x_{21C} & \dots & x_{2,p-1,C} & x_{21R} & \dots & x_{2,p-1,R} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_{n1C} & \dots & x_{n,p-1,C} & x_{n1R} & \dots & x_{n,p-1,R} \end{pmatrix} \begin{pmatrix} \beta_{0C} \\ \beta_{1C} \\ \vdots \\ \beta_{p-1,C} \\ \beta_{p,C} \\ \vdots \\ \beta_{2p-2,C} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_{1C} \\ \mathcal{E}_{2C} \\ \vdots \\ \mathcal{E}_{nC} \end{pmatrix}$$

ou encore

$$Y_C = X\beta_C + \mathcal{E}_C.$$

L'estimateur des moindres carrés de β_C est donné par

$$\boxed{\hat{\beta}_C = (X'X)^{-1}X'Y_C.}$$

On estime le centre de $y_k = [a_k, b_k]$ par

$$\hat{y}_{kC} = x_k'\hat{\beta}_C$$

où $x_k' = (1, \frac{c_{k,1}+d_{k,1}}{2}, ..., \frac{c_{k,p-1}+d_{k,p-1}}{2}, d_{k,1} - c_{k,1}, ..., d_{k,p-1} - c_{k,p-1})$.

Second modèle

Le modèle est le suivant :

$$\begin{aligned} Y_R = {} & \beta_{0R} + \beta_{1R}X_{1C} + ... + \beta_{p-1,R}X_{p-1,C} \\ & + \beta_{p,R}X_{1R} + ... + \beta_{2p-2,R}X_{p-1,R} + \mathcal{E}_R. \end{aligned}$$

Puisqu'on a mesuré Y et $X_1, ..., X_{p-1}$ sur $E = \{1, ..., n\}$, on obtient ici n relations linéaires

$$\begin{aligned} y_{iR} = {} & \beta_{0R} + \beta_{1R}x_{i1C} + ... + \beta_{p-1,R}x_{i,p-1,C} \\ & + \beta_{p,R}x_{i1R} + ... + \beta_{2p-2,R}x_{i,p-1,R} + \varepsilon_{iR} \qquad i = 1, ..., n \end{aligned}$$

où

$$y_{iR} = Y_R(i) = b_i - a_i, \qquad \varepsilon_{iR} = \varepsilon_{iU} - \varepsilon_{iL},$$

$$x_{ijC} = x_{jC}(i) = \frac{c_{ij} + d_{ij}}{2}, \qquad x_{ijR} = x_{jR}(i) = d_{ij} - c_{ij}.$$

Ces n équations peuvent s'écrire sous forme matricielle

$$\begin{pmatrix} Y_{1R} \\ Y_{2R} \\ \vdots \\ Y_{nR} \end{pmatrix} = \begin{pmatrix} 1 & x_{11C} & \dots & x_{1,p-1,C} & x_{11R} & \dots & x_{1,p-1,R} \\ 1 & x_{21C} & \dots & x_{2,p-1,C} & x_{21R} & \dots & x_{2,p-1,R} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & x_{n1C} & \dots & x_{n,p-1,C} & x_{n1R} & \dots & x_{n,p-1,R} \end{pmatrix} \begin{pmatrix} \beta_{0R} \\ \beta_{1R} \\ \vdots \\ \beta_{p-1,R} \\ \beta_{p,R} \\ \vdots \\ \beta_{2p-2,R} \end{pmatrix} + \begin{pmatrix} \mathcal{E}_{1R} \\ \mathcal{E}_{2R} \\ \vdots \\ \mathcal{E}_{nR} \end{pmatrix}$$

ou encore

$$Y_R = X\beta_R + \mathcal{E}_R.$$

L'estimateur des moindres carrés de β_R est donné par

$$\boxed{\hat{\beta}_R = (X'X)^{-1}X'Y_R.}$$

On estime l'étendue de $y_k = [a_k, b_k]$ par

$$\hat{y}_{kR} = x_k'\hat{\beta}_R$$

où $x_k' = (1, \frac{c_{k,1}+d_{k,1}}{2}, ..., \frac{c_{k,p-1}+d_{k,p-1}}{2}, d_{k,1} - c_{k,1}, ..., d_{k,p-1} - c_{k,p-1})$.

5.3.3 Prédiction de Y pour un nouvel objet

On estime la borne inférieure de $y_k = [a_k, b_k]$ par

$$\hat{a}_k = \hat{y}_{kC} - \frac{1}{2}\hat{y}_{kR}$$

et sa borne supérieure par

$$\hat{b}_k = \hat{y}_{kC} + \frac{1}{2}\hat{y}_{kR}.$$

NB : Ces prédictions n'ont de sens que si $\hat{y}_{kR} > 0$, ce qui n'est pas forcément assuré.

5.3.4 Exemple

Appliquons les deux approches de la méthode du centre et de l'étendue aux données artificielles de l'exemple 5.2.1.

Nous considérons le modèle de régression linéaire multiple :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}.$$

Dans chacune des deux approches de la méthode, nous prédirons le poids $y_k = [a_k, b_k]$ de l'objet k dont l'âge est donné par $x_{k1} = [20, 25]$ et la taille par $x_{k2} = [165, 175]$.

Première méthode du centre et de l'étendue

Premier modèle

Nous faisons une régression linéaire classique avec les centres des intervalles.

Le modèle est donc celui que nous avions calculé précédemment en appliquant la méthode du centre (exemple 5.2.1) :

$$\boxed{\hat{Y}_C = -11.571 + 0.171\, X_{1C} + 0.429\, X_{2C}.}$$

Second modèle

Nous faisons une régression linéaire classique avec les étendues des intervalles.

La matrice de régression est alors

$$X_R = \begin{pmatrix} 1 & 4 & 8 \\ 1 & 10 & 16 \\ 1 & 8 & 10 \\ 1 & 16 & 12 \\ 1 & 12 & 24 \end{pmatrix}$$

et $y_R = (8, 10, 10, 12, 12)'$.

En calculant $\hat{\beta}_R = (X_R' X_R)^{-1} X_R' y_R$, on obtient

$$\boxed{\hat{Y}_R = 6.318 - 0.300\, X_{1R} + 0.078\, X_{2R}.}$$

Prédiction

On prédit le centre de l'intervalle $y_k = [a_k, b_k]$ en utilisant le premier modèle :

$$\hat{y}_{kC} = -11.571 + 0.171 \cdot 22.5 + 0.429 \cdot 170 = 65.21$$

et son étendue en utilisant le second modèle :

$$\hat{y}_{kR} = 6.318 + 0.300 \cdot 5 + 0.078 \cdot 10 = 8.60\,.$$

On prédit donc le poids de k par

$$\hat{y}_k = [60.91, 69.51]$$

en calculant

$$\hat{a}_k = 65.21 - \frac{1}{2}8.60 = 60.91\,;$$

$$\hat{b}_k = 65.21 + \frac{1}{2}8.60 = 69.51\,.$$

Seconde méthode du centre et de l'étendue

Premier modèle

On estime le centre d'un intervalle de Y en utilisant à la fois les centres et les étendues des intervalles des régresseurs. La matrice de régression est alors

$$X = \begin{pmatrix} 1 & 20 & 164 & 4 & 8 \\ 1 & 30 & 166 & 10 & 16 \\ 1 & 40 & 168 & 8 & 10 \\ 1 & 50 & 172 & 16 & 12 \\ 1 & 60 & 176 & 12 & 24 \end{pmatrix}.$$

En calculant $\hat{\beta}_C = (X'X)^{-1} X' y_C$, on obtient

$$\boxed{\hat{Y}_C = -15.355 + 0.139\, X_{1C} + 0.452\, X_{2C} + 0.097\, X_{1R} + 0.016\, X_{2R}.}$$

Second modèle

On estime l'étendue d'un intervalle de Y en utilisant à la fois les centres et les étendues des intervalles des régresseurs. La matrice de régression est donc la même que dans le premier modèle.

En calculant $\hat{\beta}_R = (X'X)^{-1} X' y_R$, on obtient

$$\boxed{\hat{Y}_R = 21.290 + 0.077\, X_{1C} - 0.097\, X_{2C} + 0.194\, X_{1R} + 0.032\, X_{2R}.}$$

Prédiction

En utilisant le premier modèle, on prédit

$$\begin{aligned}\hat{y}_{kC} &= -15.355 + 0.139 \cdot 22.5 + 0.452 \cdot 170 + 0.097 \cdot 5 + 0.016 \cdot 10 \\ &= 65.26.\end{aligned}$$

En utilisant le second modèle, on obtient

$$\begin{aligned}\hat{y}_{kR} &= 21.290 + 0.077 \cdot 22.5 - 0.097 \cdot 170 + 0.194 \cdot 5 + 0.032 \cdot 10 \\ &= 7.82.\end{aligned}$$

On prédit donc le poids de k par

$$\hat{y}_k = [61.35, 69.17]$$

en calculant

$$\hat{a}_k = 65.26 - \frac{1}{2}7.82 = 61.35\,;$$

$$\hat{b}_k = 65.26 + \frac{1}{2}7.82 = 69.17\,.$$

□

5.3.5 Comparaison avec la méthode du centre

Nous allons appliquer les deux approches de la méthode du centre et de l'étendue aux données réelles de l'exemple 5.2.2.

Nous considérons le modèle de régression linéaire multiple :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mathcal{E}.$$

Dans chacune des deux approches de la méthode, nous prédirons la fréquence cardiaque $y_k = [a_k, b_k]$ de l'individu k dont les pressions artérielles systolique est diastolique sont respectivement données par $x_{k1} = [118, 126]$ et $x_{k2} = [85, 110]$.

Nous comparerons ensuite ces prédictions avec celle obtenue en utilisant la méthode du centre.

Première méthode du centre et de l'étendue

Premier modèle :

$$\boxed{\hat{Y}_C = 14.165 - 0.040\, X_{1C} + 0.830\, X_{2C}.}$$

Second modèle :

$$\boxed{\hat{Y}_R = 23.718 - 0.117\, X_{1R} + 0.165\, X_{2R}.}$$

Prédiction

En utilisant le premier modèle, on estime le centre de $y_k = [a_k, b_k]$ par

$$\hat{y}_{kC} = 90.21\,.$$

On estime l'étendue de y_k à l'aide du second modèle par

$$\hat{y}_{kR} = 26.91\,.$$

On prédit donc la fréquence cardiaque de l'individu k par

$$\hat{y}_k = [76.76, 103.67]$$

en calculant

$$\hat{a}_k = 90.21 - \frac{1}{2}26.91 = 76.76\,;$$

$$\hat{b}_k = 90.21 + \frac{1}{2}26.91 = 103.67\,.$$

Seconde méthode du centre et de l'étendue

Premier modèle :

$$\boxed{\hat{Y}_C = 7.247 + 0.059\,X_{1C} + 0.673\,X_{2C} - 0.044\,X_{1R} + 0.386\,X_{2R}.}$$

Second modèle :

$$\boxed{\hat{Y}_R = 2.320 + 0.155\,X_{1C} + 0.035\,X_{2C} - 0.234\,X_{1R} + 0.258\,X_{2R}.}$$

Prédiction

En utilisant respectivement le premier et le second modèle, on estime

$$\hat{y}_{kC} = 89.36 \quad \text{et} \quad \hat{y}_{kR} = 29.22\,.$$

On prédit donc la fréquence cardiaque de l'individu k par

$$\hat{y}_k = [74.75, 103.97]$$

en calculant

$$\hat{a}_k = 89.36 - \frac{1}{2}29.22 = 74.75\,;$$

$$\hat{b}_k = 89.36 + \frac{1}{2}29.22 = 103.97\,.$$

Comparaison avec la méthode du centre

Nous avons repris dans le tableau suivant les prédictions de la fréquence cardiaque $y_k = [a_k, b_k]$ obtenues en utilisant la méthode du centre et les deux approches de la méthode du centre et de l'étendue, lorsque les pressions artérielles systolique et diastolique sont respectivement $x_{k1} = [118, 126]$ et $x_{k2} = [85, 110]$.

	centre	centre + étendue 1	centre + étendue 2
$\hat{y}_k$	[80.00, 100.43]	[76.76, 103.67]	[74.75, 103.97]

On remarque que les intervalles prédits par les deux approches de la méthodes du centre et de l'étendue sont presque identiques.

Ces deux approches semblent donc aussi performantes.

L'intervalle prédit par la méthode du centre est plus étroit.

□

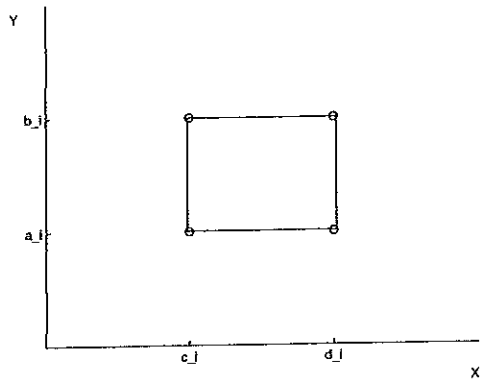
5.4 Autres méthodes

Nous proposons ici d'autres méthodes, uniquement dans le cas de la régression linéaire simple.

Ces méthodes ont été introduites par O. Rodriguez dans [Rodriguez01].

5.4.1 Les méthodes du sommet inférieur droit et du sommet supérieur gauche

Ces méthodes consistent à calculer une régression linéaire classique avec, respectivement, les sommets inférieurs droits et supérieurs gauches des rectangles obtenus en croisant les deux variables intervalles.



On prédit alors les bornes inférieure et supérieure de la variable dépendante y_k pour un objet k, respectivement par le minimum et le maximum des valeurs trouvées en appliquant ce modèle aux bornes inférieure et supérieure du régresseur x_k mesuré sur cet objet.

On peut également appliquer simultanément ces deux méthodes et estimer ensuite la borne inférieure (respectivement supérieure) de y_k par celle estimée dans la méthode du sommet inférieur droit (respectivement supérieur gauche).

NB : Notons que dans ce cas, il n'est pas sûr que la borne inférieure de l'intervalle prédit soit toujours inférieure à sa borne supérieure.

Considérons les variables *quantitatives* Y_L, X_L et Y_U,X_U qui prennent comme valeur, respectivement, les bornes inférieures et supérieures des intervalles assumés par Y et X.

Méthode du sommet inférieur droit

On estime la borne inférieure d'un intervalle de Y à partir de la borne supérieure de l'intervalle de X.

Le modèle est donc

$$Y_L = \beta_0 + \beta_1 X_U + \mathcal{E}_L.$$

Si on a mesuré X et Y sur l'ensemble $E = \{1, ..., n\}$, on obtient ici n relations linéaires

$$y_{iL} = \beta_0 + \beta_1 x_{iU} + \varepsilon_{iL} \qquad i = 1, ..., n$$

où $y_{iL} = Y_L(i) = a_i$ et $x_{iU} = X_U(i) = d_i$.

Les estimations des moindres carrés de β_0 et de β_1 sont

$$\hat{\beta}_0 = \bar{y}_L - \hat{\beta}_1 \bar{x}_U \quad \text{et} \quad \hat{\beta}_1 = \frac{s_{x_U y_L}}{s^2_{x_U}}.$$

On estime les bornes inférieure et supérieure de $y_k = [a_k, b_k]$ respectivement par

$$\hat{a}_k = \min\{\hat{\beta}_0 + \hat{\beta}_1 c_k, \hat{\beta}_0 + \hat{\beta}_1 d_k\}$$

et

$$\hat{b}_k = \max\{\hat{\beta}_0 + \hat{\beta}_1 c_k, \hat{\beta}_0 + \hat{\beta}_1 d_k\}$$

où c_k et d_k sont les bornes de $x_k = [c_k, d_k]$.

Méthode du sommet supérieur gauche

On estime la borne supérieure d'un intervalle de Y à partir de la borne inférieure de l'intervalle de X.

Le modèle est donc

$$Y_U = \beta_0 + \beta_1 X_L + \mathcal{E}_U.$$

Si on a mesuré X et Y sur l'ensemble $E = \{1, ..., n\}$, on obtient ici n relations linéaires

$$y_{iU} = \beta_0 + \beta_1 x_{iL} + \varepsilon_{iU} \qquad i = 1, ..., n$$

où $y_{iU} = Y_U(i) = b_i$ et $x_{iL} = X_L(i) = c_i$.

Les estimations des moindres carrés de β_0 et de β_1 sont

$$\hat{\beta}_0 = \bar{y}_U - \hat{\beta}_1 \bar{x}_L \quad \text{et} \quad \hat{\beta}_1 = \frac{s_{x_L y_U}}{s^2_{x_L}}.$$

On estime les bornes inférieure et supérieure de $y_k = [a_k, b_k]$ respectivement par

$$\hat{a}_k = \min\{\hat{\beta}_0 + \hat{\beta}_1 c_k, \hat{\beta}_0 + \hat{\beta}_1 d_k\}$$

et

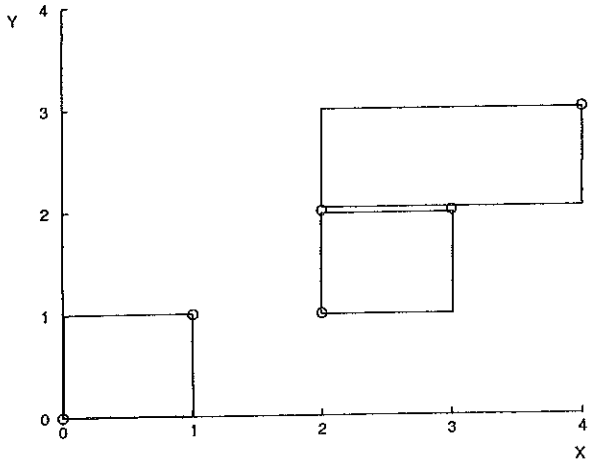
$$\hat{b}_k = \max\{\hat{\beta}_0 + \hat{\beta}_1 c_k, \hat{\beta}_0 + \hat{\beta}_1 d_k\}$$

où c_k et d_k sont les bornes de $x_k = [c_k, d_k]$.

Exemple

Considérons le tableau de données (imaginé) suivant :

k	X	Y
1	$[0,1]$	$[0,1]$
2	$[2,3]$	$[1,2]$
3	$[2,4]$	$[2,3]$



Méthode du sommet inférieur droit

On considère le tableau classique suivant :

k	X_U	Y_L
1	1	0
2	3	1
3	4	2

On calcule

$$\bar{y}_L = 1; \qquad \bar{x}_U = \frac{8}{3};$$

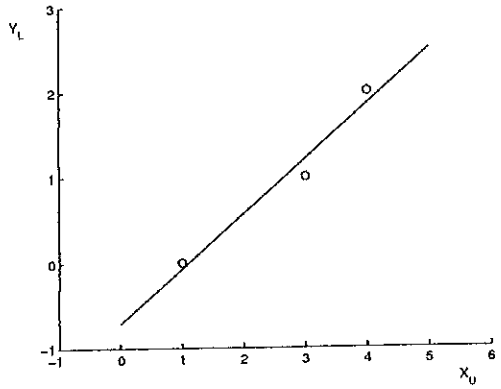
$$s^2_{x_U} = \frac{14}{9}; \qquad s_{x_U y_L} = 1.$$

On obtient alors

$$\hat{\beta}_0 = -\frac{5}{7} \quad \text{et} \quad \hat{\beta}_1 = \frac{9}{14}.$$

Le modèle obtenu est donc

$$\boxed{\hat{Y}_L = -\tfrac{5}{7} + \tfrac{9}{14} X_U.}$$



Méthode du sommet supérieur gauche

On considère le tableau classique suivant :

k	X_L	Y_U
1	0	1
2	2	2
3	2	3

On calcule

$$\bar{y}_U = 2; \qquad \bar{x}_L = \frac{4}{3};$$

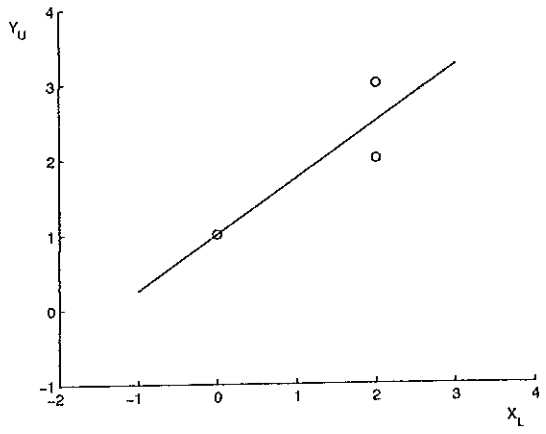
$$s^2_{x_L} = \frac{8}{9}; \qquad s_{x_L y_U} = \frac{2}{3}.$$

On obtient alors

$$\hat{\beta}_0 = 1 \quad \text{et} \quad \hat{\beta}_1 = \frac{3}{4}.$$

Le modèle obtenu est donc

$$\boxed{\hat{Y}_U = 1 + \frac{3}{4} X_L.}$$



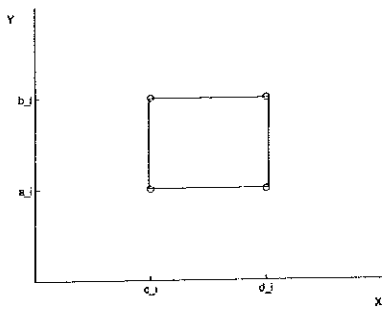
□

Remarque

Dans le cas de la régression linéaire simple, la méthode de la borne inférieure (respectivement supérieure) revient à calculer une régression classique avec les sommets inférieurs gauches (respectivement supérieurs droits) des rectangles obtenus en croisant les deux variables intervalles.

5.4.2 Régression linéaire simple avec tous les sommets

Cette méthode consiste à calculer une régression linéaire classique avec tous les sommets des rectangles obtenus en croisant les deux variables intervalles.



Plus précisément, nous calculons une régression sur base des $4 \times n$ relations linéaires

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \qquad j = 1, ..., 4n$$

où

$$y_j = \begin{cases} a_i & j = 1, ..., n \quad \text{et} \quad j = 2n+1, ..., 3n\,, \\ b_i & j = n+1, ..., 2n \quad \text{et} \quad j = 3n+1, ..., 4n\,; \end{cases}$$

$$x_j = \begin{cases} c_i & j = 1, ..., n \quad \text{et} \quad j = 3n+1, ..., 4n\,, \\ d_i & j = n+1, ..., 2n,\ 2n+1, ..., 3n\,; \end{cases}$$

$$\varepsilon_j = \begin{cases} \varepsilon_{iL} & j = 1, ..., n \quad \text{et} \quad j = 2n+1, ..., 3n\,, \\ \varepsilon_{iU} & j = n+1, ..., 2n \quad \text{et} \quad j = 3n+1, ..., 4n\,; \end{cases}$$

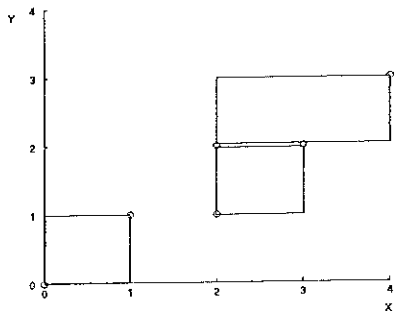
avec $i = 1, ..., n$.

On prédit alors les bornes inférieure et supérieure de la variable dépendante y_k pour un objet k, respectivement par le minimum et le maximum des valeurs trouvées en appliquant ce modèle aux bornes inférieure et supérieure du régresseur x_k mesuré sur cet objet.

Exemple

Considérons à nouveau le tableau de données

k	X	Y
1	$[0,1]$	$[0,1]$
2	$[2,3]$	$[1,2]$
3	$[2,4]$	$[2,3]$



Nous calculons une régression linéaire classique avec $4 \times 3 = 12$ individus pour lesquels les valeurs de X et de Y sont reprises dans le tableau classique suivant :

k	X	Y	k	X	Y
1	0	0	7	1	0
2	2	1	8	3	1
3	2	2	9	4	2
4	1	1	10	0	1
5	3	2	11	2	2
6	4	3	12	2	3

On calcule

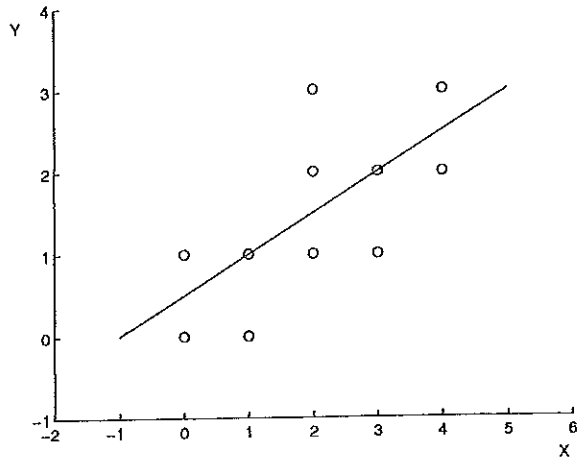
$$\bar{y} = \frac{3}{2}; \qquad \bar{x} = 2;$$
$$s_x^2 = \frac{5}{3}; \qquad s_{xy} = \frac{5}{6}.$$

On obtient alors

$$\hat{\beta}_0 = \frac{1}{2} \quad \text{et} \quad \hat{\beta}_1 = \frac{1}{2}.$$

Le modèle obtenu est donc

$$\boxed{\hat{Y} = \tfrac{1}{2} + \tfrac{1}{2} X \ .}$$



□

5.4.3 Comparaison avec la méthode du centre

Nous reprenons l'exemple mentionné par L. Billard et E. Diday dans [Billard02a].

Considérons les variables

Y : hématocrite (pourcentage du volume du sang occupé par des cellules rouges),
X : quantité totale d'hémoglobine dans le sang en grammes / décilitre.

Le tableau ci-dessous contient les valeurs (intervalles) de ces deux variables pour 16 objets :

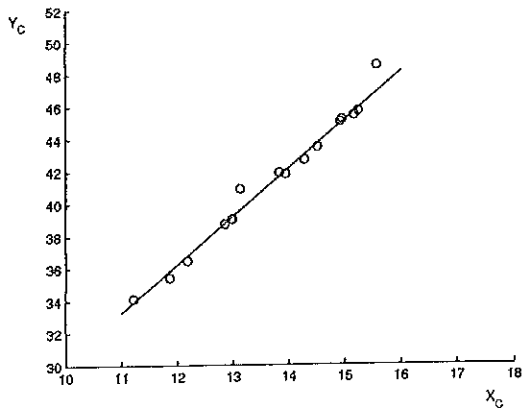
k	Y : hématocrite	X : hémoglobine	k	Y : hématocrite	X : hémoglobine
1	[32.296, 39.601]	[11.545, 12.806]	9	[28.831, 41.980]	[9.922, 13.801]
2	[36.694, 45.123]	[12.075, 14.177]	10	[44.481, 52.536]	[15.374, 15.755]
3	[36.699, 48.685]	[12.384, 16.169]	11	[27.713, 40.499]	[9.722, 12.712]
4	[36.386, 47.412]	[12.354, 15.298]	12	[34.405, 43.027]	[11.767, 13.936]
5	[39.190, 50.866]	[13.581, 16.242]	13	[30.919, 47.091]	[10.812, 15.142]
6	[39.701, 47.246]	[13.819, 15.203]	14	[39.351, 51.510]	[13.761, 16.562]
7	[41.560, 48.814]	[14.341, 15.554]	15	[41.710, 49.678]	[14.698, 15.769]
8	[38.404, 45.228]	[13.274, 14.601]	16	[35.674, 42.382]	[12.448, 13.519]

Nous avons appliqué à ces données les différentes méthodes de régression linéaire simple vues dans cette section, ainsi que la méthode du centre.

Nous avons obtenu les modèles suivants :

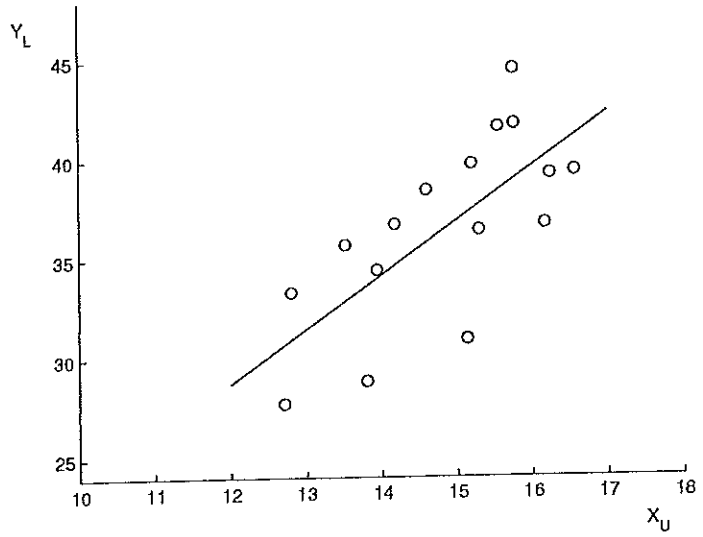
Modèle (1), méthode du centre :

$$\boxed{\hat{Y}_C = 0.497 + 2.978\, X_C}$$



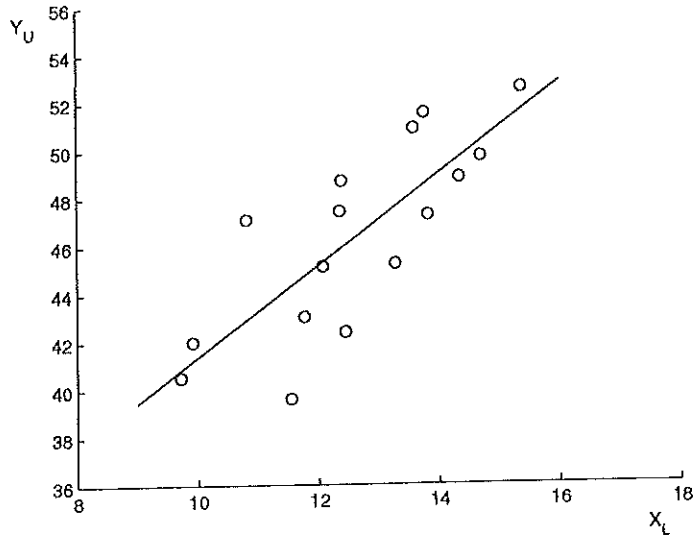
Modèle (2), méthode du sommet inférieur droit :

$$\boxed{\hat{Y}_L = -3.832 + 2.713\, X_U}$$



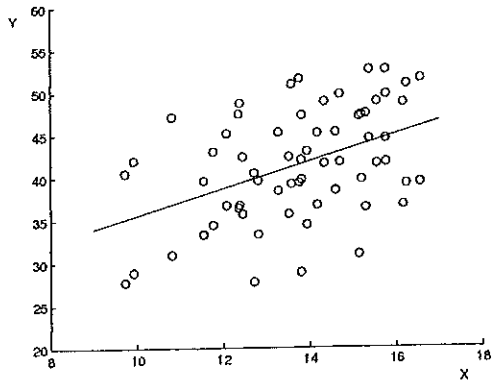
Modèle (3), méthode du sommet supérieur gauche :

$$\boxed{\hat{Y}_U = 22.262 + 1.909\, X_L}$$



Modèle (4), régression linéaire simple avec tous les sommets :

$$\boxed{\hat{Y} = 19.847 + 1.571\,X}$$



Nous avons ensuite utilisé ces différents modèles pour prédire l'hématocrite de l'objet k, $y_k = [a_k, b_k]$, lorsque la quantité d'hémoglobine dans son sang est donnée par $x_k = [12, 13]$.

Les différentes prédictions obtenues sont reprises dans le tableau suivant :

modèle	(1)	(2)	(3)	(4)
$\hat{y}_k$	$[36.23, 39.21]$	$[28.72, 31.44]$	$[45.17, 47.08]$	$[38.70, 40.27]$

Nous remarquons que l'intervalle prédit par la méthode (4) (régression avec tous les sommets) est presque identique à celui prédit par la méthode du centre.

Notons aussi que la borne inférieure de l'intervalle prédit par la méthode du centre est encadrée par les bornes inférieures des intervalles prédits par les modèles (2) et (3) :

$$28.72 < 36.23 < 45.17.$$

Il en est de même en considérant les bornes supérieures de ces intervalles :

$$31.44 < 39.21 < 47.08.$$

Lorsqu'on applique simultanément les méthodes (2) (sommet inférieur droit) et (3) (sommet supérieur gauche), on peut estimer $\hat{y}_k = [28.72, 47.08]$.

Cet intervalle est alors beaucoup plus large que celui prédit par la méthode du centre, $[36.23, 39.21]$.

La méthode du centre semble donc préférable aux autres.

□

Chapitre 6

Evaluation de la qualité de la régression et validation

6.1 Evaluation de la qualité de la régression et validation

Le coefficient de détermination, les tests d'hypothèses sur les coefficients de la régression, ainsi les intervalles de confiance pour ces coefficients et pour les prédictions, n'ont pas été rigoureusement étendus à la régression linéaire symbolique.

Cependant, une approche serait d'utiliser leurs analogues classiques, mais en prenant les centres des valeurs intervalles.

Par exemple, le coefficient de détermination R est égal au coefficient de corrélation entre Y et $\hat{Y}$:

$$R = r_{y\hat{y}} = \frac{s_{y\hat{y}}}{s_y s_{\hat{y}}}$$

où, dans le cas de variables intervalles Y et $\hat{Y}$, on a

$$s_y = \sqrt{\frac{1}{4n}\sum_{i=1}^{n}(a_i + b_i)^2 - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(a_i + b_i)\right]^2};$$

$$s_{\hat{y}} = \sqrt{\frac{1}{4n}\sum_{i=1}^{n}(\hat{a}_i + \hat{b}_i)^2 - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(\hat{a}_i + \hat{b}_i)\right]^2};$$

$$s_{y\hat{y}} = \frac{1}{4n}\sum_{i=1}^{n}(a_i + b_i)(\hat{a}_i + \hat{b}_i) - \frac{1}{4n^2}\left[\sum_{i=1}^{n}(a_i + b_i)\right]\left[\sum_{i=1}^{n}(\hat{a}_i + \hat{b}_i)\right]$$

avec $y_i = [a_i, b_i]$ et $\hat{y}_i = [\hat{a}_i, \hat{b}_i]$, $i = 1, ..., n$.

Exemple 6.1.1

Reprenons l' exemple 5.2.2 concernant la fréquence cardiaque Y en fonction des pressions artérielles systolique X_1 et diastolique X_2. Le R^2 correspondant au modèle établi par la méthode du centre est $R^2 = 0.673$.

Si nous considérons maintenant l'exemple de la section 5.4.3 concernant l'hématocrite Y en fonction de la quantité d'hémoglobine X, le R^2 du modèle obtenu en appliquant la méthode du centre est $R^2 = 0.981$.

□

Les statistiques des tests de Fisher et de Student sont également calculées en considérant les centres des intervalles.

NB : Ces valeurs sont calculées dans le logiciel SODAS 2.

Remarque

Une toute autre démarche serait d'utiliser des méthodes de validation croisée.

Par exemple, pour chaque $k = 1, ..., n$, la méthode du leave-one-out consiste à construire le modèle de régression linéaire à partir des données concernant tous les objets sauf k. Ce modèle est alors utilisé pour prédire y_k par y_k^*, étant donné les valeurs x_{kj} $(j = 1, ..., p-1)$ des régresseurs pour cet objet.

L'ensemble des valeurs ainsi prédites $\{y_k^*, k = 1, ..., n\}$ est ensuite comparé à l'ensemble des valeurs $\{y_k, k = 1, ..., n\}$ dont nous disposons.

6.2 Remarque sur la méthode du centre et de l'étendue

Sur base des indicateurs suivants

$$\sqrt{\frac{SSE_L}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_{iL} - \hat{y}_{iL})^2}{n}},$$

$$\sqrt{\frac{SSE_U}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(y_{iU} - \hat{y}_{iU})^2}{n}},$$

$$R_L = \frac{s_{y_L\hat{y}_L}}{s_{y_L}s_{\hat{y}_L}} \quad \text{et} \quad R_U = \frac{s_{y_U\hat{y}_U}}{s_{y_U}s_{\hat{y}_U}},$$

F. A. T. De Carvalho, E. A. Lima Neto et C. P. Tenorio ont montré dans [DeCarvalho04b] que, quelle que soit l'approche, la méthode du centre et de l'étendue était plus performante que la méthode du centre.

La supériorité de cette méthode par rapport à la méthode du centre augmente lorsque le nombre de régresseurs présents dans le modèle augmente.

De plus, plus la relation entre la variable dépendante et les variables indépendantes est linéaire, plus la différence entre les méthodes est significative.

Aussi, plus l'étendue des intervalles s'approche de zéro, plus la différence entre les méthodes diminue. C'est un résultat attendu car plus l'étendue des intervalles s'approche de zéro, plus la méthode du centre devient un cas particulier de la méthode du centre et de l'étendue.

Enfin, quel que soit le nombre de régresseurs présents dans le modèle, les deux approches de la méthode du centre et de l'étendue sont aussi performantes.
On va donc préférer la première approche car le nombre de paramètres à estimer dans la seconde approche est presque le double du nombre de paramètres à estimer dans la première.

Exemple 6.2.1

Considérons à nouveau les données l'exemple 5.2.2, auxquelles ont été appliquées la méthode du centre et les deux approches de la méthode du centre et de l'étendue.

Nous avons calculé les indicateurs définis ci-dessus pour les modèles obtenus en utilisant ces différentes méthodes :

méthode	R_L	R_U	$\sqrt{\frac{SSE_L}{n}}$	$\sqrt{\frac{SSE_U}{n}}$
centre	0.638	0.826	10.760	8.563
centre+étendue 1	0.725	0.818	9.109	8.275
centre+étendue 2	0.754	0.842	8.590	7.649

La supériorité des méthodes du centre et de l'étendue par rapport à la méthode du centre, établie sur base de nos indicateurs, peut donc se vérifier dans le cas de cet exemple.

□

Quatrième partie

La régression linéaire pour des données histogrammes et diagrammes

Chapitre 7

Données histogrammes

Nous présentons ici l'extension de la méthode du centre aux variables histogrammes proposée par L. Billard et E. Diday dans [Billard02a].

NB : Cette méthode a été programmée dans le logiciel SODAS 2.

Nous considérons le modèle de régression linéaire simple, la méthode s'étendant directement au cas où il y a plusieurs régresseurs. On souhaite expliquer une variable Y *de type histogramme* à l'aide d'une variable X *de type histogramme* par une relation linéaire

$$Y = \beta_0 + \beta_1 X + \mathcal{E}.$$

Supposons que ces variables ont été mesurées sur $E = \{1, ..., n\}$:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, ..., n$$

où pour tout $i = 1, ..., n$:

$$y_i = Y(i) = \{q_{i1}[a_{i1}, b_{i1}], ..., q_{it_i}[a_{it_i}, b_{it_i}]\},$$

$$x_i = X(i) = \{p_{i1}[c_{i1}, d_{i1}], ..., p_{is_i}[c_{is_i}, d_{is_i}]\}$$

avec

$$\sum_{k=1}^{t_i} q_{ik} = 1 \quad \text{et} \quad \sum_{l=1}^{s_i} p_{il} = 1.$$

Pour tout $i = 1, ..., n$, on suppose l'uniformité de la distribution à l'intérieur des intervalles $[a_{ik}, b_{ik}]$, $k = 1, ..., t_i$ et $[c_{il}, d_{il}]$, $l = 1, ..., s_i$.

On estime β_0 et β_1 par

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

et

$$\boxed{\hat{\beta}_1 = \frac{s_{xy}}{s_x^2}}$$

où les moyennes $\bar{y}$ et $\bar{x}$, et la covariance s_{xy}, ont été définies dans le chapitre 2 pour des données histogrammes :

$$\bar{y} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{k=1}^{t_i} (a_{ik} + b_{ik}) q_{ik},$$

$$\bar{x} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{l=1}^{s_i} (c_{il} + d_{il}) p_{il},$$

$$s_{xy} = \frac{1}{4n} \sum_{i=1}^{n} \left\{ \sum_{l=1}^{s_i} \sum_{k=1}^{t_i} (c_{il} + d_{il})(a_{ik} + b_{ik}) p_{il} q_{ik} \right\} - \frac{1}{4n^2} \left[\sum_{i=1}^{n} \left\{ \sum_{l=1}^{s_i} (c_{il} + d_{il}) p_{il} \right\} \right] \left[\sum_{i=1}^{n} \left\{ \sum_{k=1}^{t_i} (a_{ik} + b_{ik}) q_{ik} \right\} \right]$$

et la variance s_x^2 est

$$s_x^2 = \frac{1}{4n} \sum_{i=1}^{n} \left[\sum_{l=1}^{s_i} (c_{il} + d_{il}) p_{il} \right]^2 - \frac{1}{4n^2} \left[\sum_{i=1}^{n} \left\{ \sum_{l=1}^{s_i} (c_{il} + d_{il}) p_{il} \right\} \right]^2 .$$

NB : Nous faisons donc une régression linéaire simple classique en remplaçant, pour tout $i = 1, ..., n$, les histogrammes

$$y_i = Y(i) = \{q_{i1}[a_{i1}, b_{i1}], ..., q_{it_i}[a_{it_i}, b_{it_i}]\}$$

et

$$x_i = X(i) = \{p_{i1}[c_{i1}, d_{i1}], ..., p_{is_i}[c_{is_i}, d_{is_i}]\}$$

respectivement par les valeurs

$$\sum_{k=1}^{t_i} \frac{a_{ik} + b_{ik}}{2} q_{ik} \qquad \text{et} \qquad \sum_{l=1}^{s_i} \frac{c_{il} + d_{il}}{2} p_{il}.$$

Si on dispose de l'histogramme du régresseur X pour un nouvel objet noté 0, on peut prédire Y pour cet objet par l'intervalle $\hat{y}_0 = [\hat{a}_0, \hat{b}_0]$ où

$$\hat{a}_0 = \min \left\{ \hat{\beta}_0 + \hat{\beta}_1 \left[\sum_{l=1}^{s_0} c_{0l} p_{0l} \right], \hat{\beta}_0 + \hat{\beta}_1 \left[\sum_{l=1}^{s_0} d_{0l} p_{0l} \right] \right\},$$

$$\hat{b}_0 = \max \left\{ \hat{\beta}_0 + \hat{\beta}_1 \left[\sum_{l=1}^{s_0} c_{0l} p_{0l} \right], \hat{\beta}_0 + \hat{\beta}_1 \left[\sum_{l=1}^{s_0} d_{0l} p_{0l} \right] \right\}$$

avec $x_0 = \{p_{01}[c_{01}, d_{01}], ..., p_{0s_0}[c_{0s_0}, d_{0s_0}]\}$.

Remarque

Les variables intervalles sont des cas particuliers de variables histogrammes. En effet, lorsque $t_i = 1$, $q_{i1} = 1$ et $s_i = 1$, $p_{i1} = 1$ pour tout $i = 1, ..., n$, les histogrammes $Y(i)$ et $X(i)$ deviennent des intervalles.

Chapitre 8

Données diagrammes

Nous exposons ici la méthode de régression linéaire avec des variables explicatives diagrammes présentée par F. Afonso dans sa thèse [Afonso05a] sur base d'un exemple et explicitée dans [Afonso05b].

Cette méthode est disponible dans le logiciel SODAS 2.

On souhaite expliquer une variable Y *classique quantitative* à l'aide de $p-1$ régresseurs $X_1, ..., X_{p-1}$ *de type diagramme* par une relation linéaire

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \mathcal{E}.$$

Supposons que ces variables ont été mesurées sur $E = \{1, ..., n\}$:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{i,p-1} + \varepsilon_i \qquad i = 1, ..., n$$

où

$$y_i = Y(i) \quad \text{et} \quad x_{ij} = X_j(i) = \left\{ p_{ij1}\xi_{ij1}, ..., p_{ijs_{ij}}\xi_{ijs_{ij}} \right\}$$

avec

$$\sum_{l_j=1}^{s_{ij}} p_{ijl_j} = 1$$

pour $i = 1, ..., n; j = 1, ..., p-1$.

Pour tout $j = 1, ..., p-1$, supposons que X_j prenne s_j modalités différentes sur $E = \{1, ..., n\}$ et notons ces modalités $\xi_{j1}, ..., \xi_{js_j}$.

Nous considérons alors les variables *quantitatives* $X_{jl}, l = 1, ..., s_j$, qui prennent comme valeur, respectivement, la fréquence relative de la modalité ξ_{jl} de l'ensemble des modalités pondérées assumé par X_j.

Plus formellement, pour tout $j = 1, ..., p-1$, $l = 1, ..., s_j$ et $i = 1, ..., n$:

$$X_{jl}(i) = \begin{cases} p_{ijl_j} & \text{si } \xi_{jl} = \xi_{ijl_j} \quad l_j = 1, ..., s_{ij}; \\ 0 & \text{sinon.} \end{cases}$$

Pour chaque régresseur X_j, nous choisissons $(s_j - 1)$ des s_j variables X_{jl} $(l = 1, ..., s_j)$ lui correspondant afin que $(X'X)$ soit inversible et que nous puissions calculer le vecteur des coefficients de la régression $\hat{\beta} = (X'X)^{-1}X'y$.

Nous faisons alors une régression linéaire classique avec ces $\sum_{j=1}^{p-1}(s_j - 1)$ variables explicatives.

NB : La prédiction de Y pour un nouvel objet étant donné les diagrammes des régresseurs pour cet objet sera alors un réel.

Exemple

Nous avons imaginé l'exemple suivant afin d'illustrer la méthode présentée ci-dessus.

Considérons la variable classique continue Y : "prix" d'espace d'observations $\mathcal{Y} = \mathbb{R}^+$ et la variable diagramme X : "qualité" dont l'espace d'observations $\mathcal{X} = \{\text{mauvaise, moyenne, bonne}\}$.

Supposons que les valeurs de ces variables sur 6 objets soient les suivantes :

k	X	Y
1	{0.75 mauvaise, 0.25 moyenne}	1000
2	{0.60 mauvaise, 0.40 moyenne}	1500
3	{0.10 mauvaise, 0.80 moyenne, 0.10 bonne}	2000
4	{0.15 moyenne, 0.85 bonne}	3000
5	{0.10 moyenne, 0.90 bonne}	4000
6	{0.05 moyenne, 0.95 bonne}	4500

Considérons les variables X_1, X_2 et X_3 correspondant respectivement aux modalités mauvaise, moyenne et bonne de la variable X et prenant comme valeurs les fréquences observées pour X.

On obtient donc le tableau suivant :

k	X_1	X_2	X_3	Y
1	0.75	0.25	0	1000
2	0.60	0.40	0	1500
3	0.10	0.80	0.10	2000
4	0	0.15	0.85	3000
5	0	0.10	0.90	4000
6	0	0.05	0.95	4500

Nous calculons alors une régression linéaire classique avec 2 des variables X_i $(i = 1, ..., 3)$. Nous avons choisi X_2 et X_3.

En calculant $\hat{\beta} = (X'X)^{-1}X'y$ avec

$$X = \begin{pmatrix} 1 & 0.25 & 0 \\ 1 & 0.40 & 0 \\ 1 & 0.80 & 0.10 \\ 1 & 0.15 & 0.85 \\ 1 & 0.10 & 0.90 \\ 1 & 0.05 & 0.95 \end{pmatrix},$$

nous avons obtenu le modèle suivant :

$$\boxed{\hat{Y} = 949.3 + 871.3\,X_2 + 3135.5\,X_3.}$$

Si un objet k est de mauvaise qualité avec une probabilité de 15% et de qualité moyenne avec une probabilité de 85%, nous estimons son prix par

$$\hat{y}_k = 949.3 + 871.3 \cdot 0.85 + 3135.5 \cdot 0 = 1689.9.$$

□

Remarque

Les variables classiques qualitatives sont des cas particuliers de variables diagrammes. En effet, lorsque $s_{ij} = 1$, $p_{ij1} = 1$ $(j = 1, ..., p-1)$ pour tout $i = 1, ..., n$, les diagrammes $X_j(i)$ deviennent des catégories.

Les variables multi-catégoriques sont également des cas particuliers de variables diagrammes. En effet, si $p_{ij1} = p_{ij2} = ... = p_{ijs_{ij}} = \frac{1}{s_{ij}}$ $(j = 1, ..., p-1)$ pour tout $i = 1, ..., n$, les diagrammes $X_j(i)$ deviennent les ensembles de catégories

$$\{\xi_{ij1}, ..., \xi_{ijs_{ij}}\} \qquad j = 1, ..., p-1; i = 1, ..., n.$$

Cinquième partie

Applications

Le module SREG du logiciel SODAS 2

Un programme de régression linéaire symbolique (SREG) a été implanté. Ce programme a été inclus en tant que module du logiciel d'analyse de données symboliques SODAS (Symbolic Official Data Analysis System) développé dans le cadre du projet européen ASSO.

Nous présentons ici l'utilisation de ce programme à l'aide duquel ont été traitées les applications qui suivront.

1 Input

Les données d'entrée se présentent sous la forme d'une matrice classique ou symbolique provenant d'un fichier .sds du logiciel SODAS.

1.1 Choix des variables

L'utilisateur doit choisir la variable dépendante et les variables explicatives.

- La variable dépendante peut être classique quantitative, intervalle ou histogramme.

- Les variables explicatives peuvent être classiques quantitatives et qualitatives, intervalles, histogrammes, diagrammes, taxonomiques et hiérarchiques.

NB :

- Une variable taxonomique est une variable organisée en arbre exprimant plusieurs niveaux de généralité.

- Une variable hiérarchique, ou variable mère-filles, est une hiérarchie de va-

riables telle que des variables appelées "filles" n'ont de sens que si une autre variable appelée "mère" prend un ensemble de modalités bien spécifique.

1.2 Choix des paramètres

L'utilisateur doit ensuite choisir le test de nullité des paramètres (Fisher ou Student) qui sera utilisé lors de la sélection des régresseurs, ainsi que le niveau de confiance de ce test (95%, 99% ou 99.9%).

Par défaut, le test de Fisher est choisi à 95% de confiance.

NB : Le test de Student n'est possible qu'avec des variables quantitatives.

S'il y a des taxonomies, l'utilisateur doit choisir entre 4 méthodes (par décomposition, par agrégation, divisive, divisive multi-niveaux).

La méthode par défaut est la méthode divisive multi-niveaux.

2 Output

2.1 Evaluation du pouvoir explicatif de chaque régresseur

Le programme régresse chaque variable explicative séparément. Pour chaque modèle $Y = \beta_{0k} + \beta_k X_k$ $(k = 1, ..., p-1)$, le programme calcule

$$F_{obs} = \frac{\frac{SSE_{H_0} - SSE}{1}}{\frac{SSE}{n-2}} = \frac{SSR}{\frac{SSE}{n-2}}$$

et affiche la valeur trouvée ainsi que le quantile $F_{1-\alpha,1,n-2}$ de la loi de Fisher-Snedecor pour le niveau de confiance $(1-\alpha)$ choisi.

Ces deux valeurs sont alors comparées et le résultat du test est affiché : on rejette $H_0 : \beta_k = 0$ au niveau de signification α si $Fobs > F_{1-\alpha,1,n-2}$. Dans ce cas, le régresseur X_k est significatif.

Nous pouvons aussi voir les coefficients de détermination R^2 de ces $(p-1)$ modèles de régression linéaire simple.

NB : Dans le cas d'une variable explicative X_k qualitative, le modèle est

$$Y = \beta_{0k} + \beta_{k1} X_{k1} + ... + \beta_{km} X_{km}$$

où $X_{k1}, ..., X_{km}$ représentent m des $(m+1)$ modalités de X_k.

Le test de Fisher correspondant est celui de l'hypothèse nulle

$$H_0 : \beta_{k1} = ... = \beta_{km} = 0.$$

2.2 Sélection des variables explicatives et calcul du modèle

- Si l'utilisateur a choisi le test de Fisher, le programme calcule et affiche une régression linéaire avec les variables ayant réussi ce test.

- Si l'utilisateur a choisi le test de Student, la régression linéaire avec l'ensemble des variables est calculé et affichée.

Pour chaque régresseur, le programme nous donne

$$t_{obs} = \frac{\hat{\beta}_k}{s_{\hat{\beta}_k}}$$

et compare cette valeur au quantile $t_{1-\frac{\alpha}{2},n-p}$ de la loi de Student pour le niveau de confiance $(1-\alpha)$ choisi. Nous pouvons voir le résultat du test : on rejette $H_0 : \beta_k = 0$ au niveau de signification α si $|t_{obs}| > |t_{1-\frac{\alpha}{2},n-p}|$. Dans ce cas, la variable explicative X_k est significative.

Le programme nous donne ensuite le modèle avec les régresseurs qui ont passé ce test.

2.3 Evaluation et validation du modèle retenu

Pour le modèle retenu (q régresseurs sélectionnés), le programme calcule

$$F_{obs} = \frac{\frac{SSE_{H_0} - SSE}{q}}{\frac{SSE}{n-q-1}} = \frac{\frac{SSR}{q}}{\frac{SSE}{n-q-1}}$$

et affiche cette valeur ainsi que les quantiles de la loi de Fisher $F(q, n-q-1)$ d'ordre 95%, 99% et 99.9%.

Les résultats du test sont affichés pour ces différents niveaux de confiance : on rejette H_0 : $\beta_1 = ... = \beta_q = 0$ au niveau de signification α si $F_{obs} > F_{1-\alpha,q,n-q-1}$.

Le programme nous donne également le coefficient de détermination R^2 et les valeurs suivantes :

- sommes des carrés des écarts expliqués par la régression (SSR), résiduels (SSE) et totaux (SST) ;
- nombres de degrés de liberté :

$$DFR = q, \qquad DFE = n - q - 1, \qquad DFT = n - 1;$$

- carrés moyens :

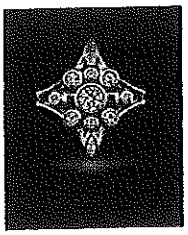
$$MSR = \frac{SSR}{DFR}, \qquad MSE = \frac{SSE}{DFE}, \qquad MST = \frac{SST}{DFT}.$$

Application 1 : diamants

1 Présentation des données

Nous souhaitons établir un modèle de régression linéaire pour expliquer le prix des diamants.

Introduction



Les facteurs qui influencent le prix d'un diamant sont les "4 C's" : carat, clarté, couleur et coupe.

- Le poids d'un diamant est indiqué en carats (1 carat = 0.2 gramme).
- Plus un diamant est parfait, plus sa clarté est grande. Les indicateurs suivants caractérisent la perfection d'un diamant : IF (Internally Flawless), VVS1 et VVS2 (Very Very Slightly Imperfect), VS1 et VS2 (Very Slightly Imperfect).
- Les indicateurs D, E, F, G, H, I caractérisent de façon décroissante la pureté de couleur d'un diamant.
- La brillance d'un diamant dépend aussi de la façon dont il a été coupé, façonné par l'artisan.

Les diamants sont évalués sur base de ces critères par différentes organisations spécialisées : le GIA (Gemmological Institue of America) de New-York, et les IGI (International Gemmological Institue) et HRD (Hoge Raad Voor Diamant) d'Anvers.

Leur réputation peut aussi être un facteur influençant le prix d'un diamant.

Données [diamants]

Les données dont nous disposons sont issues d'une publicité apparue dans le "Business Times" de Singapour du 18 février 2000. Celle-ci mentionnait, pour chaque pierre précieuse, son prix ainsi que l'organisation et son évaluation du poids, de la clarté et de la couleur.

Nous nous intéressons aux données concernant 308 diamants ronds (les autres formes moins populaires étant coeur, poire, princesse, marquise et émeraude) décrits par

- 2 variables classiques continues : prix et carat ;
- 3 variables classiques qualitatives : clarté, couleur et certif (qui correspond à l'organisation qui a évalué le diamant).

NB : Dans notre jeu de données, le prix est exprimé en dollars de Singapour (1 SGD $\simeq$ 0.50 euro) et le poids en centièmes de carat.

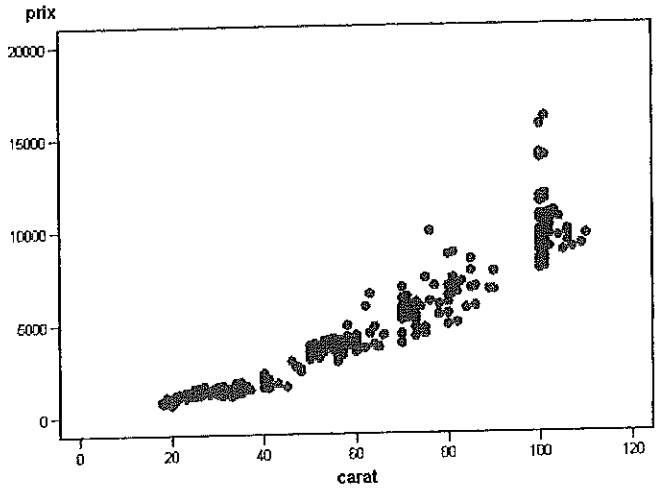
2 Régression linéaire

Nous souhaitons expliquer le prix des diamants en fonction de nos 4 régresseurs (carat, clarté, couleur et certif).

2.1 Ajustement des données

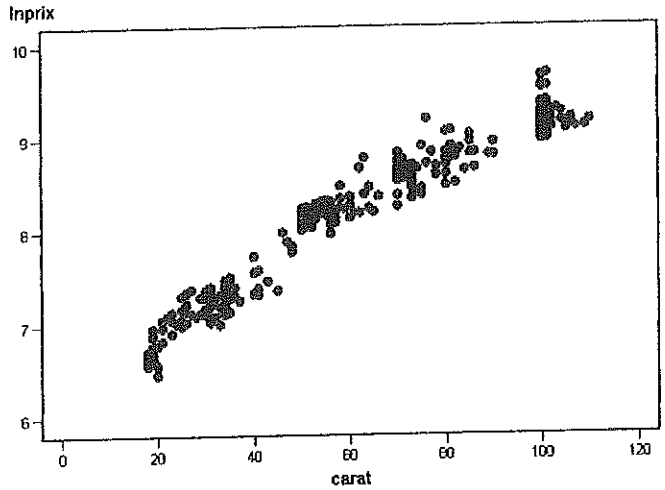
On s'attend à ce que le prix d'un diamant augmente avec son poids. Cependant, il se peut que la relation ne soit pas linéaire, les diamants plus lourds étant beaucoup plus précieux que les plus légers [Chu01].

Nous avons alors représenté le prix en fonction du poids en centièmes de carat des diamants :



On remarque que le prix des pierres plus lourdes (plus particulièrement celles pour lesquelles carat ≥ 100) augmente beaucoup plus rapidement.

Une transformation recommandée dans ce cas est de prendre le logarithme du prix. On obtient alors le graphique suivant :



Nous décidons donc d'établir le modèle de régression linéaire avec ln(prix) plutôt que prix.

Nous avons employé le module SREG de SODAS 2. Le test de nullité des paramètres utilisé est obligatoirement celui de Fisher puisque nous avons des variables qualitatives. Nous avons choisi le niveau de confiance 95%.

2.2 Régression linéaire classique

Calcul du modèle

Le programme calcule d'abord les régressions linéaires simples de chacune des variables carat, clarté, couleur et certif sur la variable dépendante ln(prix).

Nous présentons les valeurs données par SREG dans le tableau suivant :

régresseur	F_{obs}	$F_{0.95}$	R^2
carat	4441.07	4.17	0.94
certif	94.06	3.32	0.38
clarté	12.66	2.69	0.14
couleur	0.98	2.53	0.002

Nous remarquons que le test de Fisher ne rejette pas les régressions linéaires simples avec les variables carat, certif et clarté.

La variable couleur est rejetée.

La régression est ensuite calculée avec les variables sélectionnées suite au test de Fisher. Le modèle obtenu est le suivant :

$$\begin{aligned}\ln(\text{prix}) = {} & 6.595 + 0.027 \cdot (\text{carat}) - 0.165 \cdot (\text{IGI}) + 0.006 \cdot (\text{GIA}) \\ & + 0.065 \cdot (\text{VVS1}) - 0.047 \cdot (\text{VVS2}) - 0.100 \cdot (\text{VS1}) - 0.235 \cdot (\text{VS2})\end{aligned}$$

$R^2 = 0.95$ et ce modèle n'est pas rejeté par le test de Fisher aux niveaux de confiance 95%, 99% et 99.9% :

F_{obs}	$F_{0.95}$	$F_{0.99}$	$F_{0.999}$
810.23	2.33	3.31	4.82

Remarques sur le modèle

Le modèle est excellent au vu du R^2 et d'après le test de Fisher. Notons cependant que les hypothèses de normalité et d'homoscédasticité des erreurs n'ont pas été vérifiées.

Les coefficients de l'équation sont raisonnables. En particulier, les coefficients des variables indicatrices VVS1, VVS2, VS1 et VS2 reflètent bien la hiérarchie de ces grades de perfection du diamant.

2.3 Régression linéaire symbolique

Introduction

La régression symbolique est utilisée pour étudier non plus des individus du premier ordre, mais des classes d'individus appelées concepts. En effet, en agrégeant les données au niveau de ces concepts, on obtient des données symboliques.

Les concepts dépendent des centres d'intérêt de celui qui fait l'étude ou proviennent directement de la définition des données.

Dans notre application, nous étudions la variable dépendante prix. Nous avons alors réduit la taille des données en créant les concepts à partir de cette variable. Ces concepts sont en fait des intervalles de prix.

Plus précisément, nous avons construit 14 objets symboliques décrits par les 14 concepts $[90, 1500]$, $[1501, 2500]$, $[2501, 3500]$, $[3501, 4500]$, ..., $[11501, 12500]$, $[12501, 13500]$, $[13501, 16008]$.

Après la création de ces concepts, la variable classique continue carat devient également une variable intervalle et les variables classiques qualitatives clarté, couleur et certif deviennent des variables diagrammes (grâce au module DB2SO de SODAS).

Nous présentons un extrait des données classiques initiales et de leur transformation en données symboliques :

tableau classique :

k	concept	prix	carat	couleur	clarté	certif
1	[8501, 9500]	9169	100	G	$VS2$	GIA
2	[8501, 9500]	8781	105	I	$VVS2$	GIA
3	[8501, 9500]	9107	109	I	$VVS2$	HRD
4	[9501, 10500]	9713	101	G	$VS1$	IGI
5	[9501, 10500]	9890	106	H	$VVS2$	HRD

tableau symbolique :

OS	concept	prix	carat	couleur	clarté	certif
1	[8501, 9500]	[8501, 9500]	[100, 109]	$\frac{1}{3}G, \frac{2}{3}I$	$\frac{1}{3}VS2, \frac{2}{3}VVS2$	$\frac{2}{3}GIA, \frac{1}{3}HRD$
2	[9501, 10500]	[9501, 10500]	[101, 106]	$\frac{1}{2}G, \frac{1}{2}H$	$\frac{1}{2}VS1, \frac{1}{2}VVS2$	$\frac{1}{2}IGI, \frac{1}{2}HRD$

Calcul du modèle

Comme précédemment, nous étudions plutôt la régression avec la variable dépendante ln(prix).

Nous présentons le premier tableau donné par le programme SREG :

régresseur	F_{obs}	$F_{0.95}$	R^2
carat	326.80	4.75	0.96
certif	25.78	3.98	0.82
clarté	2.66	3.63	0.54
couleur	0.91	3.69	0.36

Comme dans le cas classique, les régressions linéaires symboliques simples avec les variables carat et certif ne sont pas rejetées, alors que la régression avec la variable couleur est rejetée.

Cependant, la régression symbolique avec la variable clarté est aussi rejetée, alors que la régression classique avec cette même variable ne l'était pas.

Remarque

F. Afonso fait remarquer dans sa thèse [Afonso05a] qu'en présence d'un nombre considérable d'individus dans la régression, le test de Fisher a tendance à ne jamais rejeter les régressions. Après réduction des données classiques en données symboliques, nous n'avons plus ce même problème.

Notons aussi que les R^2 de ces modèles de régression linéaire symbolique simple sont plus importants que ceux des modèles classiques correspondant.

La régression symbolique calculée avec les variables retenues suite au test de Fisher donne le modèle suivant :

$$\ln(\text{prix}) = 6.605 + 0.027 \cdot (\text{carat}) - 0.656 \cdot (\text{IGI}) + 0.210 \cdot (\text{GIA})$$

Les coefficients de cette équation semblent raisonnables.

L'ajustement linéaire de ce modèle est encore meilleur que celui du modèle classique retenu : $R^2 = 0.98$.

De plus, ce modèle n'est pas rejeté aux niveaux de confiance 95%, 99% et 99.9% :

F_{obs}	$F_{0.95}$	$F_{0.99}$	$F_{0.999}$
147.541	3.71	6.55	12.55

□

Application 2 : voitures

1 Présentation des données

Les données [voitures] dont nous disposons sont issues de "Kiplinger's Personal Finance" et datent de décembre 2003. Il s'agit de la description de différents types de véhicules (voitures, camionnettes et camions).

Nous nous intéressons ici aux données concernant 384 véhicules décrits par 4 variables classiques quantitatives :

- puissance (en chevaux),
- nombre de cylindres,
- longueur (en pouces ; 1 pouce $\simeq$ 2.5 cm),
- prix (en U.S. Dollars).

2 Régression linéaire

Nous souhaitons établir un modèle de régression linéaire expliquant la puissance d'un véhicule en fonction de son nombre de cylindres, de sa longueur et de son prix.

Nous avons utilisé le module SREG de SODAS 2 en choisissant le test de Student au niveau de confiance 95% pour la sélection des régresseurs.

2.1 Régression linéaire classique

Le programme nous donne le premier tableau suivant :

régresseur	F_{obs}	$F_{0.95}$	R^2
prix	792.06	4.17	0.67
longueur	84.67	4.17	0.18
cylindre	592.78	4.17	0.61

Les régressions linéaires simples avec chacune des variables explicatives prix, longueur et cylindre ne sont donc pas rejetées par le test de Fisher.

La régression linéaire avec l'ensemble de nos 3 régresseurs est ensuite calculée. Nous reprenons les valeurs données par SREG dans le tableau suivant :

	coeff.	t_{obs}	résultat Student
cste.	-8.125	-0.342	coeff. $= 0$
prix	0.002	16.174	
longueur	0.290	1.940	coeff. $= 0$
cylindre	16.580	9.221	

Nous remarquons que le test de Student ne rejette pas les hypothèses de nullité de la constante et du coefficient de la variable longueur.

La régression linéaire est alors calculée avec les régresseurs retenus suite au test de Student, c'est-à-dire avec les variables prix et cylindre :

$$\text{puissance} = 0.002 \cdot (\text{prix}) + 25.384 \cdot (\text{cylindre}).$$

Les coefficients de cette équation sont en accord avec le sens commun.

$R^2 = 0.79$ et ce modèle n'est pas rejeté par le test de Fisher aux niveaux de confiance 95%, 99% et 99.9% :

F_{obs}	$F_{0.95}$	$F_{0.99}$	$F_{0.999}$
725.554	3.32	5.39	8.77

2.2 Régression linéaire symbolique

Introduction

Nous avons réduit la taille des données de 384 individus à 11 objets symboliques en créant 11 concepts à partir de la variable dépendante puissance. Ces concepts sont les 11 intervalles de puissance $[73, 119]$, $[120, 139]$, $[140, 159]$, $[160, 189]$, ..., $[260, 279]$, $[280, 299]$, $[300, 500]$.

Après la création de ces concepts, les variables classiques quantitatives cylindre, longueur et prix deviennent également des variables intervalles.

Calcul du modèle

Nous présentons le premier tableau donné par le programme SREG :

régresseur	F_{obs}	$F_{0.95}$	R^2
prix	114.32	5.12	0.93
longueur	20.33	5.12	0.69
cylindre	12.09	5.12	0.57

Nous remarquons que, comme dans le cas classique, le test de Fisher ne rejette aucune des régressions linéaires avec une seule des variables prix, longueur ou cylindre.

La régression linéaire symbolique est ensuite calculée avec l'ensemble de ces variables explicatives. Le programme nous donne les résultats suivants :

	coeff.	t_{obs}	résultat Student
cste.	21.15	0.078	coeff. $= 0$
prix	0.005	4.702	
longueur	0.140	0.075	coeff. $= 0$
cylindre	0.004	0.0003	coeff. $= 0$

Comme dans le cas classique, les hypothèses de nullité de la constante et du coefficient de la variable longueur ne sont pas rejetées par le test de Student.

L'hypothèse d'un coefficient nul pour la variable cylindre n'est pas rejetée non plus, alors qu'elle l'était dans le cas classique.

NB : La sélection d'une unique variable explicative reflète bien la forte corrélation entre les 3 régresseurs proposés.

La régression symbolique est ensuite calculée avec la seule variable non rejetée suite au test de Student, c'est-à-dire avec "prix" :

$$\text{puissance} = 0.006 \cdot (\text{prix}).$$

Cette équation semble raisonnable.

Le coefficient de détermination de ce modèle est $R^2 = 0.92$. Il est donc plus élevé que celui du modèle classique retenu (0.79).

Le programme nous donne aussi les valeurs suivantes pour ce dernier modèle de régression linéaire symbolique :

F_{obs}	$F_{0.95}$	$F_{0.99}$	$F_{0.999}$
111.002	4.97	10.04	21.04

Ce modèle n'est donc pas rejeté par le test de Fisher.

□

Conclusion

L'objectif de ce mémoire était d'étendre la régression linéaire classique aux données symboliques.

Dans le cas de données intervalles, nous avons étudié différentes méthodes. Parmi ces méthodes, nous avons constaté que les méthodes du centre et de l'étendue et la méthode du centre semblaient être les plus performantes.

Nous avons aussi présenté l'extension de la méthode du centre de L. Billard et E. Diday aux données histogrammes, ainsi que la méthode de régression linéaire avec des variables explicatives diagrammes proposée par F. Afonso.

Le cas des variables taxonomiques et hiérarchiques n'a pas été traité dans ce mémoire. Des méthodes de régression linéaire ont cependant également été développées pour ces types de variable.

Ces méthodes de régression linéaire symbolique permettent l'analyse de concepts.

Dans nos applications concernant des jeux de données de taille assez importante, nous avons remarqué que la régression linéaire au niveau de concepts définis à partir de la variable dépendante donnait des résultats intéressants par rapport à l'analyse classique des individus du premier ordre. Notamment, les tests de Fisher et de Student sont moins efficaces en présence d'un grand nombre d'individus dans la régression.

Cependant, ces tests et le coefficient de détermination n'ont pas été étendus de façon rigoureuse aux données symboliques. Il serait important de poursuivre les recherches dans ce sens.

Bibliographie

[Afonso02] F. Afonso: " Régression Linéaire sur Données Symboliques : Définitions Théoriques de Méthodes pour la Régression dans le cas de Données Taxonomiques et de Variables Mères-Filles et Construction d'un Module de Régression Symbolique pour le Logiciel SODAS".
Mémoire de Stage en Analyse des Données Symboliques, Université de Paris, Dauphine, 2002.

[Afonso04] F. Afonso, L. Billard et E. Diday : "Symbolic Linear Regression with Taxonomies",
in *Studies in Classification, Data Analysis and Knowledge Organization: Classification, Clustering and Data Mining Applications*, D. Banks, L. House, F.R. McMorris, P. Arabie et W. Gaul (eds.), Springer, 2004, pp. 429-437.

[Afonso05a] F. Afonso: "Méthodes Prédictives par Extraction de Règles en Présence de Données Symboliques ".
Thèse de Doctorat, Université de Paris, Dauphine, 2005.

[Afonso05b] F. Afonso, L. Billard et E. Diday: "Symbolic Linear Regression in the Presence of Taxonomy and Hierarchical Variables ".
Technical Report, 2005.

[Bertrand00] P. Bertrand et F. Goupil : "Descriptive Statistics for Symbolic Data",
in *Analysis of Symbolic Data : Exploratory Methods for Extracting Statistical Information from Complex Data*, H.-H. Bock et E. Diday (eds.), Springer, 2000, chap. 6, pp. 106-124.

[Billard00] L. Billard et E. Diday : "Regression Analysis for Interval-Valued Data",
in *Data Analysis, Classification and related Methods: Proceedings of the Seventh Conference of the International Federation of Classification Societies, IFCS-2000*, H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen et M. Schader (eds.), Springer, 2000, pp. 369-374.

[Billard02a] L. Billard et E. Diday: "Symbolic Regression Analysis",
in *Classification, Clustering and Data Analysis: Proceedings of the Eighenth Conference of the International Federation of Classification Societies, IFCS-2002*, K. Jajuga, A. Sokolowski et H.-H. Bock (eds.), Springer, 2002, pp. 281-288.

[Billard02b] L. Billard et E. Diday : " Symbolic Data Analysis: Definitions and Examples". Technical Report, <http://www.stat.uga.edu/faculty/LYNNE/Lynne.html>, 2002.

[Billard03] L. Billard et E. Diday : "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis",
in *Journal of the American Statistical Association*, vol. 98, 2003, pp. 470-487.

[Bock00] H.-H. Bock: "Symbolic Data",
in *Analysis of Symbolic Data : Exploratory Methods for Extracting Statistical Information from Complex Data*, H.-H. Bock et E. Diday (eds.), Springer, 2000, chap. 3, pp. 39-53.

[Ceremade03] Laboratoire du Lise-Ceremade : "SODAS",
<http://www.ceremade.dauphine.fr/~touati/sodas-pagegarde.htm>,
Université de Paris, Dauphine, 2003.

[Chu01] S. Chu: "Pricing the C's of Diamond Stones",
in *Journal of Statistics Education*, vol. 9, num. 2, 2001.

[Dauphine04a] Université de Paris, Dauphine: "SREG Help Guide: Regression",
in *Help Guide of the SODAS 2 Software*, FUNDP and ASSO Scientific Partners (eds.),
<http://www.info.fundp.ac.be/asso/sodaslink.htm>, 2004.

[Dauphine04b] Université de Paris, Dauphine: "Symbolic Linear Regression by Using the Module SREG: a Short Tutorial for Users",
in *User Manual of the SODAS 2 Software*, FUNDP and ASSO Scientific Partners (eds.),
<http://www.info.fundp.ac.be/asso/sodaslink.htm>, 2004.

[DeCarvalho04a] F.A.T. De Carvalho, E.A. Lima Neto et C.P. Tenorio : "A New Method to Fit a Linear Regression Model for Interval-Valued Data",
in *Advances in Artificial Intelligence : Proceedings of the 27th German Conference on Artificial Intelligence, KI-2004. Lecture Notes on Artificial Intelligence, LNAI-3238*,
S. Biundo, T. Frühwirth et G. Palm (eds.), Springer, 2004, pp. 295-306.

[DeCarvalho04b] F.A.T. De Carvalho, E.A. Lima Neto et C.P. Tenorio : "Univariate and Multivariate Linear Regression Methods to Predict Interval-Valued Features",
in *Proceedings of the 17th Australian Joint Conference on Artificial Intelligence, AI-2004. Lecture Notes on Artificial Intelligence, LNAI-3339*, G.I. Webb et X. Yu (eds.), Springer, 2004, pp.526-537.

[Hardy03] A. Hardy : "Statistiques", Librairie des Sciences, FUNDP, Namur, 2003.

[Hardy04] A. Hardy: "Modèles Statistiques Linéaires", FUNDP, Namur, 2004.

[Raju97] S.R.K. Raju : "Symbolic Data Analysis in Cardiology",
in *Symbolic Data Analysis and its Applications*, E. Diday et K.C. Gowda (eds.), 1997,
pp. 245-249.

[Rodriguez01] O. Rodriguez: "Classification et Modèles Linéaires en Analyse des Données Symboliques".
Thèse de Doctorat, Université de Paris, Dauphine, 2001.

Jeux de données utilisés dans les applications

[diamants] <http://www.amstat.org/publications/jse/v9n2/4c1.dat>.

[voitures] <http://www/amstat.org/publications/jse/datasets/04cars.dat>.